PROSODIC AIDS TO SPEECH RECOGNITION

IV. A GENERAL STRATEGY FOR PROSODICALLY-GUIDED SPEECH UNDERSTANDING

SPERRY UNIVAC

PREPARED FOR

ADVANCED RESEARCH PROJECTS AGENCY

29 MARCH 1974

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Univac Defense Systems Division
P. O. Box 3525
St. Paul, Minnesota 55165

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
Prosodic Aids to Speech Recognition IV. A General Strategy for Prosodically-Guided Speech Understanding

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Semiannual Technical Report; 1 September, 1973 - 31 March, 1974

5. AUTHOR(S) (First name, middle initial, last name)
Wayne A. Lea

6. REPORT DATE
29 March 1974

7a. TOTAL NO. OF PAGES
76

7b. NO. OF REFS
47

8a. CONTRACT OR GRANT NO.
DAHC15-73-C-0310

b. PROJECT NO.

9a. ORIGINATOR'S REPORT NUMBER(S)
Univac Report No. PX 10791

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
None

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

13. ABSTRACT

A strategy is outlined for acoustic aspects of speech recognition, whereby prosodic features are used to detect boundaries between phrases, then stressed syllables are located within each constituent, and a partial distinctive features analysis is done within stressed syllables. Analysis of phonetic recognition results by several research groups showed that automatic phone categorization is much more accurate in stressed syllables. Prosodic features appear to be potentially useful for providing cues to sentence type, syntactic bracketing, occurrences of coordination and subordination, and specific semantic structures. Preliminary acoustic analyses are then followed by an analysis-by-synthesis strategy.

Studies showed that the time intervals between stressed vowels in several recorded texts tended to cluster at around 400 to 500 milliseconds, but the number of intervening unstressed syllables had a much more prominent effect on interstress interval than might have been expected from published hypotheses. Several interpretations of the notion that English is a stress-timed language appear to be refuted by these results. Further studies of rhythm and rate of speech are to be conducted.

Preliminary studies of some sentences that gave problems to speech understanding systems showed that prosodies do differ in yes/no questions versus commands, and that ambiguous syntactic structures can be disambiguated from prosodic patterns. A set of speech texts are being designed for careful analysis of the effects on prosodic patterns due to various contrasts in syntactic structure, semantics, stress patterns, and phonetic sequences.

Unclassified
Security Classification

14

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech Recognition						
Speech Analysis						
Linguistic Stress						
Prosodies						
Prosodic Features Extraction						
Syntactic Boundary Detection						
Distinctive Features Estimation						
Syntactic Analysis						
Syntactic Parsing						
Rhythm						

ia

SPERRY UNIVAC
COMPUTER SYSTEMS

PROSODIC AIDS TO SPEECH RECOGNITION: IV. A GENERAL STRATEGY FOR PROSODICALLY-GUIDED SPEECH UNDERSTANDING

by

Wayne A. Lea

Defense Systems Division
St. Paul, Minnesota
(612) 456-2434

Semiannual Technical Report Submitted To:

Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

Attention: Director, IPT

29 March 1974

Report No. PX10791

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC 15-73-C-0310, ARPA Order No. 2010. The views and conclusions contained in this document are those of the author, and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

PREFACE

This is the fourth in a series of reports on Prosodic Aids to Speech Recognition. The first report, subtitled ''I. Basic Algorithms and Stress Studies'', appeared 1 October 1972, as Univac Report No. PX 7940. (The subtitle did not appear on all copies of that report.) The second report, suntitled ''II. Syntactic Segmentation and Stressed Syllable Location'', appeared 15 April 1973, as Univac Report No. PX 10232. The third report, subtitled '' III. Relationships Between Stress and Phonemic Recognition Results'', appeared 21 September 1973, as Univac Report No. PX 10430.

This research was supported by the Advanced Research Projects Agency of the Department of Defense, under Contract No. DAHC15-73-C-0310, ARPA Order No. 2010. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U. S. Government.

SUMMARY

Sperry Univac is continuing its implementation and testing of a strategy of speech recognition, whereby certain acoustic features (called "prosodic features") are used to segment the speech into grammatical phrases and to identify those syllables that are given prominence, or stress, in the sentence structure. Then, partial distinguishing features analysis is to be done within each stressed syllable and wherever else reliable segmental analysis can be readily accomplished. Positions of the boundaries between grammatical phrases, stressed syllable locations, pauses and specific rhythmic patterns, and special intonational features are to be used to guide the selection of a candidate grammatical structure for the spoken sentence. This preliminary hypothesis about the grammatical structure of the sentence is thus made from prosodic features only, without prior determination of the words in the sentence. In essence, then, prosodic features are used to guide the efficient determination of the vowels and consonants making up the words in the sentence, and the determination of some aspects of the grammatical context in which those words are found.

From the partial distinguishing features analysis, and some knowledge of contextual constraints, words can be hypothesized as occurring at specific places in the utterance, with particular emphasis given to the hypothesizing of important words centered around the stressed syllables. These hypothesized words, the acceptable grammatical structures, the semantic constraints, and knowledge of the limited types of things that can be said in any specific task situation all provide the information needed to make a total hypothesis as to the identity of the spoken sentence. By an analysis-by-synthesis procedure, such an hypothesis can be submitted to sound structure rules which generate a comprehensive sound structure ("acoustic phonetic") pattern which can be compared to the sound structure pattern of the input. If the generated pattern is very similar to the input pattern, the hypothesized sentence structure is asserted to be the identity of the input sentence. If the patterns differ substantially, an error signal is produced, to guide the selection of another structural hypothesis to try.

Not all aspects of this general prosodically-guided analysis-by-synthesis strategy are being implemented for ARPA, but the strategy provides a framework within which substantial contributions to speech understanding can be provided by the judicious

use of prosodic features. This approach to speech understanding is motivated by several factors. For one thing, there is considerable evidence that human listeners make initial decisions about large linguistic units (phrases) before they attempt to decide upon the complete phonetic structures (that is, the sequence of vowels and consonants) in an utterance. They appear to use prosodic patterns such as intonation, stress, pauses, and rhythm to guide their decisions about large linguistic units.

In addition, some of the most reliable information about intended vowels and consonants is to be found in the stressed syllables, which are more carefully articulated. A study of five different methods for automatic classification of segments of speech into specific vowel and consonant categories showed that, with any of the available methods, the categorization of vowels, stop consonants (p, t, k, b, d, g), and fricatives (s, z, f, v, etc.) was far more reliable in the stressed syllables than elsewhere in the utterances. This demonstrates that stressed syllables provide "islands of reliability" in the sound structure of spoken English. These stressed syllables also occur in the most semantically important words of a sentence.

Since stressed syllables do provide some of the most important and reliably decoded information in the speech wave a procedure for locating stressed syllables in connected speech has been developed. This procedure uses acoustic parameters of energy, syllabic duration, and voice fundamental frequency ("pitch") to locate the vowel and semi-vowel sounds forming the "nucleus" of a stressed syllable. Computer programs have been implemented for classifying the stressed vowel into one of five categories, depending upon the natural vocal tract resonances (formants) to determine whether the speaker's tongue is high or low, front or back, or retroflexed in his mouth.

Automatic detection of a few particular consonant categories is also attempted, independent of the locations of stressed syllables. These detected phone categories include sibilants [s, z, ʃ , ʒ , tʃ , dʒ], r-like sounds [r, ɝ], unvoiced stops [p, t, k], and nasals [m, n, ŋ]. Some, but not all, of the occurrences of these consonants will be within stressed syllables. Then, the remaining portions of stressed syllables, not found within the stressed vowel, or within one of the detected consonantal segments, are classified into gross left-over categories of unvoiced consonant, voiced consonant, or silence. The unstressed syllables are not totally segmented, and thus there are isolated islands of detected sound structure, some being the total stressed syllables and others being one of the four types of detected consonants.

Among the recent studies at Sperry Univac has been the comparison of three approaches to stressed syllable location. Methods based on only the duration of high energy chunks, or upon only the length of time that fundamental frequency (F_0) was not falling significantly, did not perform as well as the original algorithm based on archetype F_0 contours in phrases and local searches for high-energy chunks of speech. The archetype contour algorithm was also least sensitive to the type of sentence being processed, while the other algorithms showed quite different performance in yes/no questions.

Prosodic information can be used in several important components of speech understanding systems. Stressed syllables can form the anchors around which a search for occurrences of words can be attempted. We plan to investigate how detected positions of constituent boundaries, located stressed syllables, features of intonation, pauses, and rhythms can be used to determine: the type of sentence spoken (that is, whether or not it was a yes/no question); the correct grouping of portions of the utterance into phrases and specific linguistic units; the occurrence of coordinate structures; the subordination of one phrase under another; and the occurrence of specific semantic structures like coreference, contrast, and emphasis.

We have recently undertaken a study of rhythm in our available speech texts. These studies suggest that, while stressed syllables tend to occur at about .4 to .5 seconds apart, the time intervals between stresses are very much affected by the number of unstressed syllables between stresses. In none of our texts did more than four unstressed (or reduced) syllables ever occur between two stressed syllables. There seems to be an almost equal increment in the size of the interval between onsets of stressed vowels with each increase in the number of intervening unstressed syllables. There is some suggestion in the rhythm data that the preferred rhythmic structure is an alternation of stress and unstress.

These rhythm studies also showed that the time intervals between detected phrase boundaries tended to be integral multiples of the mean time between stressed vowels. Pauses between embedded clauses tended to be of the same duration as the mean time interval between stressed vowels, while durations of pauses between sentences tended to be about twice that interstress interval.

We plan to study interstress time intervals, intervals between detected phrase boundaries, and time intervals between all syllables (or, equivalently, number of syllables per unit time) as acoustic measures of rhythm and rate of speech. Information about rate of speech may be used in selecting the appropriate phonological rules to apply in determining underlying phonemic structure from the slurred, coarticulated phonetic sequences. "Fast speech" rules show more slurring, coarticulating, and dropping of speech sounds. In addition to such phonological use of rate of speech, specific rhythmic effects such as interruptions of rhythm (pauses, "disjunctures", etc.) could be useful in hypothesizing the grammatical structure of a sentence.

Several "problem sentences", which were quite similar in the sequence of words they were composed from, but which had different syntactic structures, were submitted to our prosodic analysis procedures. These sentences, obtained from Bolt Beranek and Newman, included a yes/no question ("Have any people done chemical analyses on this rock?") and the command that resulted when the first word was incorrectly recognized ("Give any people done chemical analyses on this rock.") The command is structurally ambiguous, with one interpretation referring to the process of giving to any people those chemical analyses that have been done, and another interpretation referring to any chemical analyses that were done by people (that is, "people-done" analyses). An example of a pronunciation with each intended interpretation was provided, along with an apparently "neutral" pronunciation which was presumably intended to not indicate which of the two interpretations was correct. Another pronunciation of the wording of the yes/no question, with more like the intonation of a declarative or command, was also provided.

A study of the prosodic patterns in these sentences gave encouraging indications that prosodies can provide important cues to syntactic structure. The auxiliary verb "have" in the yes/no question was unstressed while the command verb "give" was stressed. The general slope of the F_0 contour in the yes/no question was flatter than the more-falling contour of the command, although the F_0 rise expected at the end of the yes/no question did not occur. The word "done" was stressed except when it was in the compound construction "people-done", in which case the first syllable of the following word "chemical" was stressed. There was a phrase boundary between "people" and "done" in each command except the "people-done" interpretation, in which case the boundary occurred after "done", thus marking "people-done" as a unit. The time intervals between vowel onsets also showed distinctions between the two interpretations. The "neutral" version of the command turned out to be identified with the first ("[any people] [done chemical analyses]") interpretation by every available prosodic cue. This is in harmony with an expectation that the first interpretation is the most likely, and the most like a neutral, "unmarked" interpretation, since other possible structures, like the yes/no question, would make "any people" a unit, and not "people-done" a likely unit.

Studies with these few sentences are only suggestive of possible prosodic cues to syntactic structures. Further studies with many more utterances are needed. We are currently designing an extensive set of sentences which provide "minimal pairs" of sentences with nearly identical word sequences but contrasting structures. These sentences include explicit tests of the prosodic effects of sentence type, contrastive syntactic bracketing, subordination, coordination, syntactic categories (such as pronouns, verbals, compound nouns, etc.), movement of stress within phrases, coreference, etc. Prosodic patterns to be studied for these sentences include: performance of the program for detecting phrase boundaries from valleys in F_0 contours; acoustic correlates of stressed syllables, and performance in automatic stressed syllable location: acoustic measures of rhythm and rate of speech; overall F_0 contour shapes; and local variations in prosodic features due to phonetic sequences. Also being designed are a set of sentences which include all word-initial consonant-vowel (CV) sequences, and all word-final vowel-consonant (VC) sequences. These "phonetic-sequence sentences" provide the speech data needed for efficiently testing automatic procedures for vowel and consonant classification. For example, five sentences provide instances of all distinguishable stressed vowels of American English, coupled with the sibilants [s, ʃ], in initial CV and final VC positions.

From extensive studies with such designed sentences, we hope to develop experimentally-validated intonation rules and other prosodic rules. These rules will then be used to guide parsing, semantic analysis, phonological analyses, and word matching procedures in ARPA speech understanding systems.

TABLE OF CONTENTS

1. INTRODUCTION

This is a report on work currently in progress in the Univac Speech Communications Group, under contract with the Advanced Research Projects Agency (ARPA). As a part of ARPA's total program in research on speech understanding systems, the research reported herein is concerned with extracting reliable prosodic and distinctive features information from the acoustic waveform of connected speech (sentences and discourses). Studies are being concentrated on problems of detecting stressed syllables and syntactic boundaries, doing distinctive features analysis within stressed syllables, and using prosodic features to guide syntactic parsing and semantic analysis.

Prosodic cues to sentence structure, and prosodic aids to the location of reliable acoustic phonetic information, have been given little or no attention in previous speech recognition efforts. The strong motivations for the use of prosodic patterns in speech recognition procedures were thus presented in some detail in an earlier report (Lea, Medress, and Skinner, 1972a, section 2). Improvements in the Univac facilities for extracting prosodic features, spectral data, and formants, and a program for detecting boundaries between syntactic phrases (constituents), were described in a subsequent report (Lea, Medress, and Skinner, 1973a). Extensive experiments were also described in that report, which were conducted to: (1) determine the success of detecting boundaries between major syntactic units from fall-rise patterns in fundamental frequency contours; (2) determine listeners' abilities to perceive stressed, unstressed, and reduced syllables in read texts and spontaneous utterances; and (3) determine the success of locating stressed syllables by an algorithm which used rising fundamental frequency and high energy integral as major accoustic correlates of stressed syllables in the constituents delimited by the boundary detector.

This previous work provided abilities to detect about 90% of all major syntactic boundaries from acoustic data, to locate 85% or more of the stressed syllables in connected speech, to provide reliable results about listeners' perceptions of stress levels, and to provide basic parameterization tools such as linear prediction, formant tracking, fundamental frequency tracking, and energy contours. It was assumed that stressed syllables would provide the most reliable information about phonemic content of an utterance and thus, when good distinctive features estimation procedures were developed (presumably based on the available parameterization techniques), they would work best in the stressed syllables. Later modifications and additions to prosodic and distinctive features extraction procedures (Lea, Medress, and Skinner, 1973b) provided improved fundamental frequency tracking, two new "sonorant energy" functions, voicing decisions independent of fundamental frequency tracking, and elimination of about half of the "false alarms" in syntactic boundary detection. With techniques similar to those presented at the Carnegie-Mellon University Segmentation Workshop, significant success in vowel classification and strident fricative location was attained in some preliminary experiments.

Implementation of the stressed syllable location algorithm described in an earlier report (Lea, Medress, and Skinner, 1973a) was delayed by the implementation and testing of several alternative ways of locating stressed syllables from energy and fundamental frequency contours (Lea, Medress, and Skinner, 1973b). An additional new effort was undertaken to dramatically justify the Univac strategy of early analysis of stressed syllables, by showing that segmentation and classification of vowels and consonants in continuous speech was more accurate in stressed syllables, for each of five different segmentation and classification procedures reported at the Carnegie-Mellon University Speech Segmentation Workshop. This extensive study, which is summarized in Section 2, firmly demonstrates the validity of what has previously been a general assumption of more reliable decoding in stressed syllables.

Section 2 of this report presents a general overview of a strategy for using prosodic features to guide various critical aspects of a speech understanding system. This represents the first comprehensive summary of our general strategy, which is also briefly described in an IEEE conference paper (Lea, Medress, and Skinner, 1974). Besides presenting a general analysis-by-synthesis strategy, and a description of our present implementation of prosodically-guided phonetic analyses, Section 2 provides ideas for prosodic aids to word matching, parsing, and semantic analysis, and ways in which phonetic sequence information could alter or confirm prosodic decisions.

In Section 3, several experiments are described which relate prosodic patterns to linguistic structure and talker performance. The results of some initial studies of rhythmic patterns are presented in Section 3.2, for the Rainbow Script, the Monosyllabic Script, and the 31 ARPA test sentences. Some specific ways in which prosodic information may be useful in a speech understanding system are illustrated by the preliminary results, in Section 3.3, of a study of some BBN problem sentences. Speech texts with various minimum-pair comparisons, as described in Section 3 3, provide the data for controlled experiments on various factors influencing prosodic patterns.

Conclusions and a summary of further studies to be undertaken are presented in Section [illegible] ection 5 provides pertinent references.

2. A PROSODICALLY-GUIDED SPEECH UNDERSTANDING STRATEGY

2.1 General Approach

A traditional model of speech recognition has assumed that, by tracking the right "information-carrying" parameters, and using any of several pattern-matching phonemic-segment classification techniques, one could determine phonemic strings corresponding to these intended by the talker. Then, the phonemic strings may be applied to higher-level linguistic analyses to determine words, phrases, and utterance meanings. We have argued elsewhere (Lea, 1972, 1973b; Medress, 1972; Lea, Medress, and Skinner, 1972a) that successful understanding of spoken sentences involves *early* use of linguistic structure in combination with the *most reliable* acoustic information. Because of the structural redundancy provided by the listener's linguistic knowledge, a speaker does not have to encode into the acoustic waveform all of the features describing an utterance, and the features that he does choose to encode can vary from one repetition of a given sentence to the next. Indeed, in some utterances, whole phonemes or syllables may be "missing" from the pronounciation. A speech recognition system based on the acoustic manifestation of all phonemes or all distinctive features would thus frequently fail.

In addition to this incompleteness and variability, the encoding of phonemic information is a complex one involving overlapping acoustic features and environmental dependency. Phonological rules and accustic phonetic rules must be incorporated in systems to adequately characterize the complex process of encoding phonemic information into acoustic features. Some linguists have gone so far as to argue that syntactic cues must be used even before an adequate phonemic representation can be determined (Chomsky and Miller, 1963; Lea, 1973b). Perception theorists also have argued that decisions about large linguistic units must be made before one fills in the phonemic details about an utterance (cf. review in Lea, 1973b).

Speech understanding (by man or machine) then appears to involve making use of certain expectations and received cues to determine the syntactic structure (and semantic content) of an utterance. Given an hypothesis as to the surface syntactic structure, the perceiver uses phonological principles to determine a phonetic shape.

The hypothesis will be accepted if its associated acoustic phonetic shape isn't too radically different from the acoustic input (Chomsky and Halle, 1968, p. 24).

To make the preliminary syntactic hypotheses called for in the early stage of recognition schemes, without depending upon a complete segmental (phonemic) analysis, we have proposed the use of prosodic features to segment continuous speech into sentences and phrases and locate the prominent (or stressed) syllables in those phrases. The total framework within which such prosodic information is to be coupled with acoustic phonetic and structural information is illustrated in Figure 1. This may be recognized to be an analysis-by-synthesis system, with what is usually called the "preliminary analysis" block here broken down into: a component for extracting prosodic features (energy and voicing functions); a component for extracting phonetic parameters (spectral features and formants); a prosodic structure analysis which obtains phrase boundaries, rhythms, and stress patterns; a component for obtaining a partial representation of the phonetic segment structure (or "distinguishing features matrix") within stressed syllables and wherever other reliable phone categorizations can be accomplished; and a preliminary syntactic hypothesizer which uses phrase boundaries, rhythms, and stress patterns, together with allowable syntactic structures specified by the grammar, to predict the likely syntactic structure of the sentence. The top three boxes in Figure 1 thus provide prosodically-derived guesses as to the large-unit structure of an utterance, independent of detailed phonetic analyses. The phonetic parameter extraction and partial distinguishing features estimation are also guided by prosodic information which indicates where detailed phonetic analyses should be done.

Following such preliminary analyses, the lexical hypothesizer proposes possible lexical entries for insertion in the sentence structure, based on the closest match between the partial distinguishing features representation of the input and the lexical entries in the lexicon. Contextual constraints, such as lexical categories that could occur at certain positions in the sentence structure, and likely words in certain semantic and task contexts, are used to guide the lexical hypothesizing.

The acceptable grammatical structures dictated by the grammar, and the constraints about what can be said based on the semantic structures and task domain, combine together with the lexical and syntactic hypotheses resulting from the preliminary analysis of the input speech, to yield a total hypothesis about the identity of the

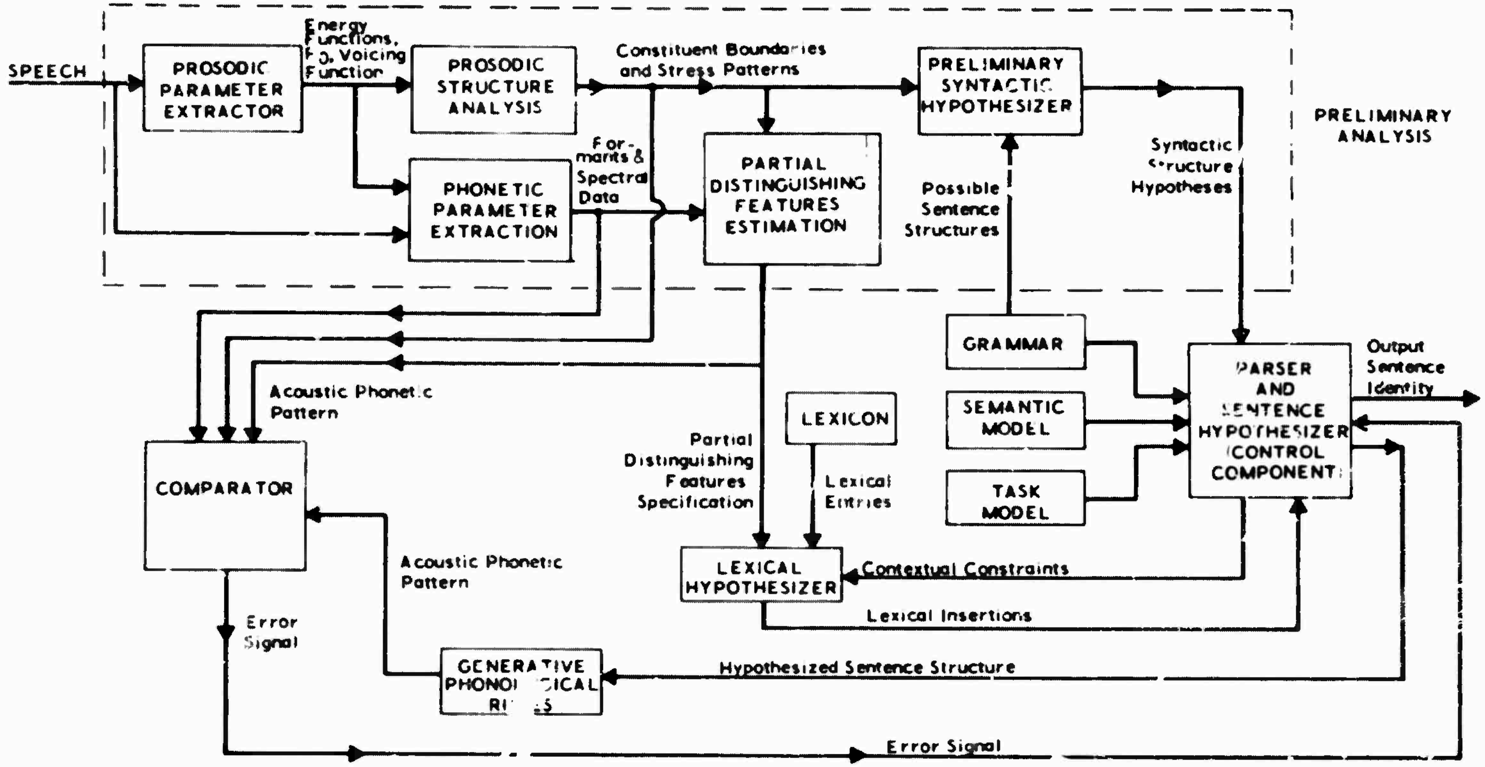

SPEECH
PROSODIC PARAMETER EXTRACTOR
Energy Functions, F0, Voicing Function
PROSODIC STRUCTURE ANALYSIS
PHONETIC PARAMETER EXTRACTION
For-mants & Spectral Data
Constituent Boundaries and Stress Patterns
PARTIAL DISTINGUISHING FEATURES ESTIMATION
PRELIMINARY SYNTACTIC HYPOTHESIZER
Possible Sentence Structures
Syntactic Structure Hypotheses
PRELIMINARY ANALYSIS
GRAMMAR
SEMANTIC MODEL
TASK MODEL
PARSER AND SENTENCE HYPOTHESIZER (CONTROL COMPONENT)
Output Sentence Identity
LEXICON
Lexical Entries
Partial Distinguishing Features Specification
LEXICAL HYPOTHESIZER
Contextual Constraints
Lexical Insertions
Hypothesized Sentence Structure
GENERATIVE PHONOLOGICAL RULES
Acoustic Phonetic Pattern
Acoustic Phonetic Pattern
COMPARATOR
Error Signal
Error Signal

Figure 1. A Prosodically-Guided Analysis-by-Synthesis System

sentence spoken. The preliminary analysis should exclude from the parser's consideration all but a small subset of possible structures to hypothesize. The sentence hypothesizer then controls the order in which hypothesized sentence structures are generated and applied to the generative phonological rules (including acoustic phonetic rules), to yield acoustic phonetic patterns for comparison with the input acoustic phonetic patterns. It is anticipated that the comparisons between input and internally generated patterns will be done at the level of distinguishing features patterns and such prosodic patterns as phrase boundaries and stress patterns, but it is possible that some comparisons can best be done at the parameter level of formants and the like. Consequently each of these three types of extracted patterns have been shown feeding to the comparator of Figure 1, to be compared with the internally generated patterns.

This overall prosodically-guided analysis-by-synthesis strategy, and the preliminary analysis routines in particular, will ultimately have substantial effects on the form of parser, semantic analysis routines, and phonological rules implemented in the system, as will be discussed further in the remainder of section 2. The grammar, lexicon, semantic model, task model, and the basic analysis-by-synthesis blocks of sentence hypothesizer, phonological rules, and comparator are vital components in this strategy, but work has not progressed to the stage where they can be implemented. We do anticipate working closely with Bolt Beranek and Newman on the use of a locally-organized parser and augmented transition network grammar, to work with the locations of reliably-encoded stressed syllables, out to the recognition of local islands of structure, which are then to be united by hypothesized structural connections. Prosodic aids to word matching, phonological analysis, parsing, and semantic analysis will be discussed in later portions of Section 2. In Section 2.2, we consider the only aspects of the prosodically-guided speech understanding strategy that have been implemented (in at least an initial form) at Sperry Univac.

2.2 Prosodic Guidelines to Phonetic Analysis

The preliminary analysis components shown within the dotted lines in Figure 1 provide all the information derived from the acoustic signal, which is to be compared with the stored and generated information in the other components. In Figure 2 is shown the version of these preliminary analysis components that is currently being

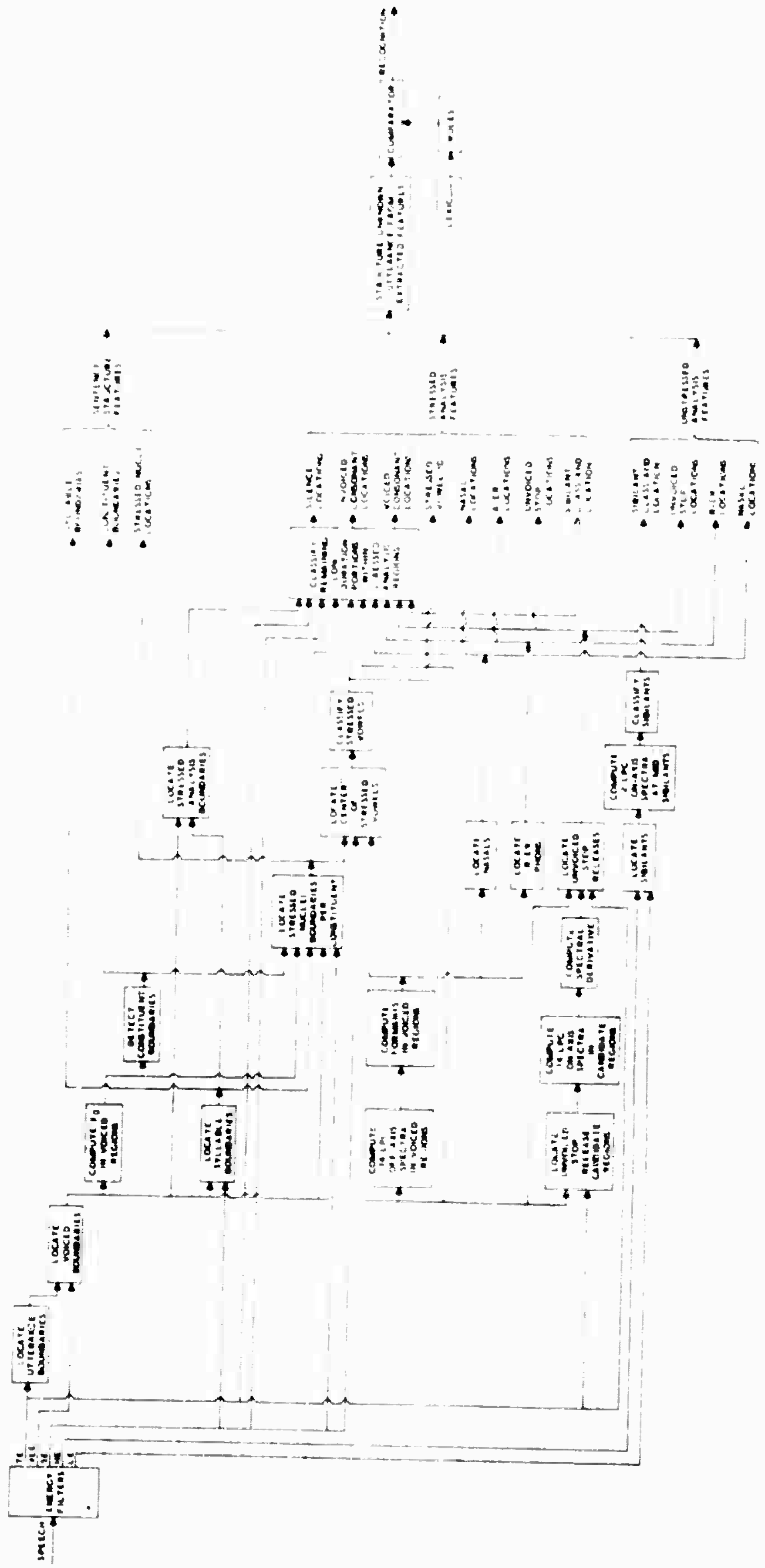

SPEECH
ENERGY FILTERS
LOCATE UTTERANCE BOUNDARIES
LOCATE VOICED BOUNDARIES
COMPUTE F0 IN VOICED REGIONS
LOCATE SYLLABLE BOUNDARIES
DETECT CONSTITUENT BOUNDARIES
LOCATE STRESSED NUCLEI BOUNDARIES PER CONSTITUENT
LOCATE STRESSED ANALYSIS BOUNDARIES
LOCATE CENTER OF STRESSED VOWELS
CLASSIFY STRESSED VOWELS
CLASSIFY REMAINING LOW DURATION PORTIONS WITHIN STRESSED ANALYSIS REGIONS
COMPUTE 14 LPC SPECTRA IN VOICED REGIONS
COMPUTE FORMANTS IN VOICED REGIONS
LOCATE UNVOICED STOP RELEASE CANDIDATE REGIONS
COMPUTE 14 LPC ON AXIS SPECTRA IN CANDIDATE REGIONS
COMPUTE SPECTRAL DERIVATIVE
LOCATE NASALS
LOCATE R L W PHONE
LOCATE UNVOICED STOP RELEASES
LOCATE SIBILANTS
COMPUTE 2 LPC SPECTRA AT MID SIBILANTS
CLASSIFY SIBILANTS
SYLLABLE BOUNDARIES
CONSTITUENT BOUNDARIES
STRESSED NUCLEI LOCATIONS
SENTENCE STRUCTURE FEATURES
SILENCE LOCATIONS
UNVOICED CONSONANT LOCATIONS
VOICED CONSONANT LOCATIONS
STRESSED VOWEL ID
NASAL LOCATIONS
R L W LOCATIONS
UNVOICED STOP LOCATIONS
SIBILANT CLASS AND LOCATION
STRESSED ANALYSIS FEATURES
UNSTRESSED ANALYSIS FEATURES
STRUCTURE UNKNOWN UTTERANCE FROM EXTRACTED FEATURES
COMPARATOR
RECOGNITION
LEXICON
WORDS

Figure 2. Preliminary Acoustic Analysis Components

implemented on the new Sperry Univac speech research facility. Here we will briefly survey the various processes and resulting parameters and decisions that result from the "acoustic front end" shown in Figure 2.

2.2.1 Prosodic Aids to Efficient Analysis – One of the advantages of the prosodically-guided preliminary analysis routine shown in Figure 2 is the potential for more efficient analysis. Prosodic features limit the places where the most complex computations of parameters must be done.

As shown in Figure 2, hardware energy filters provide functions of total energy (TE) between 60 and 5000 Hz, sonorant energy (SE) between 60 and 3000 Hz, high frequency energy (HE) between 3000 and 5000 Hz, low frequency energy (LE) between 60 and 900 Hz, and very low frequency energy (VLE) between 60 and 400 Hz. The utterance beginning and end are determined from the total energy function, and the beginnings and ends of all voiced portions in the utterance are determined from the very low frequency energy function. Only within those regions recognized as voiced are spectral analyses, formant tracking and fundamental frequency tracking done each 10 milliseconds (ms). Similarly, only in the unvoiced regions of high total energy following region of low total energy (indicating the burst following the silence of an unvoiced stop) is the spectral derivative computed, for locating unvoiced stop bursts. Phonetic identifications are only attempted where they can be readily and accurately accomplished. The sound structure in the region of the stressed syllable, and the locations of sibilants, [r]'s, unvoiced stop releases, and nasals provide some of the more reliable phonetic information needed for identifying a sentence.

Prosodic guidelines also can improve efficiencies in the higher levels of linguistic analysis. The number and positions of constituent boundaries, stressed syllables, and syllable boundaries are total-sentence-structure indications that may be used to rule out candidate sentences and direct the efficient search for the most likely sentences to compare with. Prosodies can provide some cues as to how to disambiguate syntactic and semantic structures, as we shall see later in this report.

2.2.2 Stressed Syllables as Islands of Reliability – In addition to the increased efficiency provided by prosodically-guided speech analysis, our approach to recognition focuses on the most reliable information in the speech signal. There is enough redundancy in speech, and enough evidence that listeners do large-unit recognition without prior attention to all phonetic details of the speech, to suggest that all the phonetic structure need not be given equal attention in recognition schemes (cf. Lea, 1973b). It has often been assumed that consonants and vowels should be more clearly articulated and easier to distinguish in stressed syllables than in unstressed or reduced syllables, so that a partial distinctive features analysis in the region of the stressed syllable should provide some of the most reliable phonetic information. Linguists say this is not true in some languages, such as certain dialects of Tajik and of Syrian Arabic (in which stressed vowels are not significantly longer than unstressed vowels, are actually less clear, and show fewer phonemic contrasts than unstressed vowels), or in languages like Spanish (in which stress has no important effect on the quality of vowels) [cf. Ferguson, 1966, p. 54]. However, in languages like English and Russian, stress has the effect of lengthening a vowel and enhancing its characteristic coloration, so that stressed vowels are expected to be clearer and more likely to reach target positions.

In experiments with trained subjects reading spectrograms, Klatt and Stevens (1972) have shown that partial transcriptions of vowels were correct 50% of the time for stressed vowels, 44% for unstressed vowels, and 30% for reduced vowels. Prestressed consonants were also shown to be more accurately recognized from visual study of spectrograms than were consonants in other positions (74% correct identification in the prestressed consonants compared to a maximum of 29% in any other position; Klatt and Stevens, 1972).

A recent pilot study at Sperry Univac showed than an algorithm for locating strident fricatives located 91% of the phonemically strident fricatives in stressed syllables, 86% in unstressed syllables, and 66% in reduced syllables, for 31 ARPA test sentences (listed in Lea, Medress, and Skinner, 1973b, pp. 18-20). A very preliminary approach to stop location located 46% of all phonemic stops in stressed syllables, and only about half that percentage in unstressed and reduced syllables. Two thirds of all located stops were in stressed syllables. The highest percentage of stop location was 60%, in prestressed single stops (just before stressed vowels). Higher percentages of single stops (by about 15%) were located than for stops within clusters (Lea, Medress, and Skinner 1973b).

We may expect that phonetic structure, whether it is automatically determined or determined by humans listening or reading spectrograms, will correspond more closely with underlying lexical representations (or ''phonemic'' structure) in stressed syllables than in unstressed or reduced syllables. A separate issue, however, is whether automatic categorization of the phones in continuous speech will correspond more closely with a trained listener's phonetic transcription in stressed syllables than elsewhere in the speech. Do available techniques for categorizing phonetic segments from acoustic features better match phonetic transcriptions in stressed syllables than in unstressed or reduced syllables? To answer this question, we conducted a study at Sperry Univac of the results of several approaches to automatic segmentation and labeling of continuous speech (Lea, 1973e).

Nine research groups participating in the Carnegie-Mellon Speech Segmentation Workshop segmented the 31 ARPA test sentences into phone-size portions and identified the phones, phonetic categories, or distinctive features of the segments. We shall not here consider the accuracy of placement of segment boundaries; only the correspondence between their automatic categorizations of segments (such as high/mid/low vowel, front/central/back vowel, sonorant, nasal, unvoiced obstruent, etc.) and the phonetic transcriptions will be considered. The phonetic transcription provided by a linguist (Linda Shockey) at the Carnegie-Mellon Workshop was accepted as the standard for comparison. Five groups supplied sufficiently detailed vowel and consonant categorizations to warrent their inclusion in this initial study. They shall be referred to here as Groups A, B, C, D, and E.

The sophistication and detailed methods of segment categorization used by these groups varied considerably, but for each group a chart was compiled giving a score for the level of correspondence between each of the acoustically-derived segment labels and the phonetician's labels. Figure 3 shows a sample of the type of correspondence chart used. If the phonetic transcription said a vowel was an /i/ and the group's automatic categorization called the majority of that time segment a front vowel, a score of +2 was assigned. If it was called a central vowel, then +1 was assigned, and so forth. A vowel categorized as an unvoiced stop, or the converse, would be assigned a low score of -2. Then, all categorizations in the connected speech which had a score of +1 or greater where considered acceptable, while lower scores were not acceptable. Figure 4 shows a similar chart for evaluating the categorizations by another group. This group gave more exact categorizations into phone categories. In essence, these charts are a crude attempt to characterize phonetic similarity and to assign a threshold

SEGMENT CATEGORIZATION ↓ / PHONETIC TRANSCRIPTION CATEGORY →	DIPHTHONGS	VOWELS	SONORANTS	FRICATIVES VOICED	FRICATIVES UNVOICED	AFFRICATES	STOPS VOICED	STOPS UNVOICED
	[illegible]	[illegible] u ʊ	l r [illegible] n ŋ [illegible] y	[illegible] ʒ ʒ v	[illegible] ʃ ɦ f h	dʒ tʃ	b d g	p t k
SrV	+. Ø Ø Ø→ +. Ø→ +.	[illegible] Ø Ø +1 Ø Ø Ø Ø Ø Ø	Ø Ø Ø Ø Ø Ø +1	-1 -1 -1 -1	-. -. -. -. -.	-. -.	-1 -1 -1	-. -. -.
c-V	+1 +1 +1 +1 +1	+1 +1 +1 +1 +1 +. +. +. + +1 +1 1 +1 +1	Ø +1 Ø Ø Ø Ø Ø	-1 -1 -1 -1	-. -. -. -. -	- -.	-1 -1 -1	-. -. -.
LxV	Ø +2 +. +.→Ø +.→Ø	Ø Ø Ø Ø Ø +1 +1 +1 +1 +. +. +. +. -	Ø +1 + Ø Ø Ø Ø	-1 -1 -1 -1	-. -. -. +. -.	-. -2	-1 -1 -1	-. +2 -.
N (nasal)	Ø Ø Ø Ø Ø	Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø	+1 +1 +. +. +. +1 +1	-1 -1 -1 -1	-. -. -. -. -.	-. -.	-1 -1 -1	-. +. -.
G (glide)	Ø Ø Ø Ø Ø	Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø Ø	+1 +1 +1 +1 +1 +2 +.	-1 -1 -1 -1	-. -. -. -. -.	-. -2	-1 -1 -1	-2 -.
F (fricative)	-2 -. . -. -.	- -. -. -. -. -. -. -. -. -. -2 -2 -2 -2	-1 -1 1 -1 -1 -1 -1	+1 +1 +. +2	+1 +1 +. +. +2	+1 +1	Ø Ø Ø	Ø Ø Ø
S (strident fricative)	-. -. - -2	-. -2 -2 -2 -2 -. -. -2 -. -2 -2 -2 -2	-1 -1 -1 -1 -1 -1 -1	+2 +2 +1 +1	+2 +2 +1 +1 +1	+2 +2	Ø Ø Ø	Ø Ø Ø
t-p (labial/alveolar obstruent)	-. -2 -. . -.	-. -2 -. -. -. -. -. -2 -. -. -2 -2 -2 -.	-1 -1 -1 -1 -1 -1 -1	+1 +1 +1 1	+1 +1 +2 +2 +2	+1 +1	Ø Ø Ø	+2 +1 +1
ʃ-x (velar or palatal obstruent)	-2 -2 - -2 -2	- -. -. -. -2 -2 -2 -2 -2 -2 -. -. -2 -.	-1 -1 -1 -1 -1 -1 -1	+1 +2 +1 +1	+1 +2 +1 +1 +2	+2 +2	Ø Ø Ø	+1 +1 +2
D Voiced Stop	-2 -. -. -2 -.	-. -2 -. -. -2 -2 -2 -2 -2 -2 -2 -2 -2 -2	Ø Ø Ø Ø Ø Ø Ø	+1 +1 +1 +1	Ø Ø Ø Ø Ø	+2 +1	+2 +2 +2	+1 +1 +1
- silence	-2 -. -. -2 -2	-2 -2 -2 -2 -2 -. -. -2 -2 -. -. -. -2 -2	-1 -1 -1 -1 -1 -1 -1	-1 -1 Ø Ø	1 -1 Ø Ø Ø	+1 +1	+1 +1 +1	+2 +2 +2

Figure 3. Correspondence Chart for Evaluating the Automatic Segment Categorization by Group A

PHONETIC TRANSCRIPTION CATEGORY →	DIPTHONGS					VOWELS														SONORANTS							FRICATIVES									AFFRI- CATES		STOPS					
																											VOICED				UNVOICED							VOICED			UNVOICED		
SEGMENT CATEGORIZATION ↓	aI	[illegible]	aU	eI	ɔI	i	I	ɛ	æ	[illegible]	[illegible]	ʌ	Ɨ	ɝ	[illegible]	u	ʊ	ɔ	o	l	r	m	n	ŋ	w	y	z	ʒ	ð	v	s	ʃ	θ	f	h	dʒ	tʃ	b	d ɾ	b	p	t	k
i	+1	ø	ø	+1	+1	+1	+1	+1	+1	+1	+1	+1	+2	+1	ø	+1	+1	ø	ø	ø	ø	ø	ø	ø	ø	+1	-1	1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
I	+1	ø	ø	+1	+1	+2	+1	+1	+1	+1	+1	+1	+1	+1	ø	+1	+1	ø	ø	ø	ø	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
ɛ	+1	ø	ø	ø	ø	+1	+2	+3	+2	+2	-1	+1	+1	ø	ø	ø	+1	ø	ø	ø	ø	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
æ	+1	ø	ø	+1	ø	+1	+1	+1	+3	+1	+1	+1	+1	+1	+1	ø	ø	+1	+1	ø	ø	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
ə	ø	ø	ø	ø	ø	+1	+1	+1	+1	+1	+1	+2	+1	+1	+1	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
ʌ	ø	ø	ø	ø	ø	+1	+1	+1	+1	+1	+2	+3	+1	1	+1	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
ɝ	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	+1	+1	+1	+3	ø	+1	+2	+1	+1	ø	+1	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
a	ø	+1	+3→+1	+3→ø	+2→ø	ø	ø	ø	+1	ø	+1	+1	ø	ø	+3	+1	+1	+2	+1	ø	ø	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
u	+1	+1→+2	+1→+2	+1→ø	+1→ø	+1	+1	ø	ø	ø	ø	ø	+1	+1	+1	+3	+2	+1	+1	ø	ø	ø	ø	ø	+1	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	1	-1	-2	-2	-2
ʊ	+1	+1→+2	+1→+2	+1→ø	+1→ø	+1	+1	+1	ø	+1	+1	+1	+1	+2	+1	+2	+2	+2	+2	ø	ø	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
o	+1	+3	+1→2	+1	+2→+1	ø	ø	ø	+1	+1	+1	+1	ʒ	+1	+2	+2	+2	+2	+3	ø	ø	ø	ø	ø	ø	ø	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
R	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	+2	ø	ø	ø	ø	ø	+1	+3	+1	+1	+1	+1	+1	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-2	-2
r	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	+2	ø	ø	ø	ø	ø	+1	+3	+1	+1	+1	+1	+1	-1	-1	-1	-1	-2	-2	-2	-2	-2	2	-2	-1	-1	-1	-2	-2	-2
n	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	3	ø	ø	ø	ø	ø	+1	+1	+2	+3	+2	+1	+1	-1	-1	-1	-1	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	2	-2
s	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	2	-2	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	+1	+1	+1	+1	+3	+2	+2	+2	+1	ø	+1	ø	ø	ø	ø	+1	ø
ʃ	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	+1	+1	+1	+1	+2	+3	+2	2	+1	ø	+1	ø	ø	ø	ø	ø	+1
f	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	+1	+1	+1	1	+2	+2	+3	+3	+2	ø	+1	ø	ø	ø	+1	ø	ø
h	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	+1	+1	+1	+1	+2	+2	+2	+2	+3	ø	+1	ø	ø	ø	ø	ø	ø
p	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	ø	ø	ø	ø	ø	ø	ø	+1	ø	ø	+1	+2	+1	+1	+3	+2	+2
t	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	ø	ø	ø	ø	+1	ø	+1	ø	ø	ø	+2	+1	+2	+1	+2	+3	+2
k	-2	-2	-2	-2	-2	-2	-2	-2	2-	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	ø	ø	ø	ø	ø	+1	ø	ø	ø	ø	+1	+1	+1	+2	+2	2	+3

Figure 4. Correspondence Chart for Evaluating the Automatic Segment Categorization by Group D

of acceptable similarity between phonetic transcription and acoustically-derived categories.

Obviously, there are many times when a phonetician's assignment of segment boundaries do not line up with those automatically derived. While the placement of boundaries was not explicitly studied, some judgments of sufficient proximity and overlap were needed to properly compare the automatic results with the right phonetic units. Each phonetic segment was required to overlap for at least twenty milliseconds with an acceptably-labeled automatic categorization.

The 31 test sentences were also submitted to a panel of listeners for classification of all syllables in the speech as stressed, unstressed, or reduced. Also, an algorithm for automatic location of stressed syllables was used to delimit stressed nuclei. Tables I and II show how, for segmentation results by Groups A and D, the stress level of a syllable affects the probability of a failure to provide acceptable phonetic categorizations for the vowels and consonants in that syllable.

The error rates are shown separately for syllables perceived (by a majority of listeners) as reduced, unstressed, or stressed. Obviously, vowels and obstruents are least likely to be inadequately categorized in stressed syllables. Interestingly, sonorants do not show the same form of dependence upon stress. Dependence upon stress is particularly prominent with the less sophisticated (more errorful) segmentation methods.

We may conclude that, while ideal methods might be devised to phonetically categorize as well in unstressed and reduced syllables as in stressed syllables, several available (and practical) methods for phonetic segmentation correspond most closely with phonetic transcriptions in stressed syllables. Combining this with the closer correspondence expected between phonetics and underlying phonemic structure in stressed syllables, and the semantic importance assigned to stressed syllables, one can see the value of early attention to stressed syllables in procedures for recognition of continuous speech.

Because of this demonstrated importance of stressed syllables in speech recognition, an algorithm has been devised for locating stressed syllables from prosodic features of energy and fundamental frequency (Lea, 1973a, f; Lea, Medress, and Skinner, 1972b; 1973a, b). This algorithm, which has been described in detail elsewhere (Lea, 1973a), is based on local increases in fundamental frequency, and large integrals of

TABLE I
PERCENTAGES OF ALL CATEGORIZATIONS THAT WERE UNACCEPTABLE, FOR GROUP A

	STRESSED	UNSTRESSED	REDUCED
VOWELS	12%	19%	47%
SONORANTS	42	38	30
FRICATIVES	8	16	17
STOPS	10	25	28

TABLE II
PERCENTAGES OF ALL CATEGORIZATIONS THAT WERE UNACCEPTABLE, FOR GROUP D

	STRESSED	UNSTRESSED	REDUCED
VOWELS	0%	3%	7%
SONORANTS	8	6	5
FRICATIVES	10	10	19
STOPS	6	20	29

energy within the syllabic nucleus, being the most reliable acoustic correlates of stress. It also incorporates adjustments based on the most common ("archetype") fundamental frequency contours within the grammatical phrases and clauses of connected speech.

To evaluate this algorithm for stressed syllable location, perception tests have been conducted to determine which syllables in connected speech are perceived as stressed. Extensive tests (Lea, 1973a,d; Lea, Medress, and Skinner, 1973b), with several listeners repeating, at several separate occasions, their assignments of which are the stressed, unstressed, and reduced syllables in 400 seconds of connected speech, showed that the best listeners agreed with each other, and were consistent from time to time, on 95% of their judgments about which were the stressed syllables. This thus provided a "standard" of stress assignment that was consistent within a 5% tolerance, so that, within that 5% level of precision, perceptions of stress could be compared with automatic locations of stressed syllables.

On the average, over 85% of all syllables perceived as stressed were correctly located by the archetype contour algorithm.

To further evaluate the effectiveness of this archetype contour algorithm for locating stressed syllables, these results were compared with results in stressed syllable location by other procedures. Figure 5 shows one simple procedure which finds all dips and peaks in the sonorant energy function and delimits syllabic nuclei as all contiguous points within 5 dB of the maximum intensity value in each high-intensity "chunk" or syllable. Then, those chunks (or syllabic nuclei) that have a minimum duration of 100 ms are declared to be stressed.

The results of applying this simple stressed syllable location program are shown in Table IIIa. The location of stressed syllables from durations of high-intensity chunks works surprisingly well in read texts with sharply contrasting stress levels, such as the Monosyllabic Script, but it is not as effective in more complicated read texts such as the Rainbow Script or in spontaneous speech such as the ARPA Sentences. Lowest percentages of correct location and highest percentages of false alarms occur for the spontaneous ARPA sentences.

Another simple routine, shown in Figure 6, locates all portions of speech where, for 100 ms or longer, fundamental frequency does not decrease more than one eighth

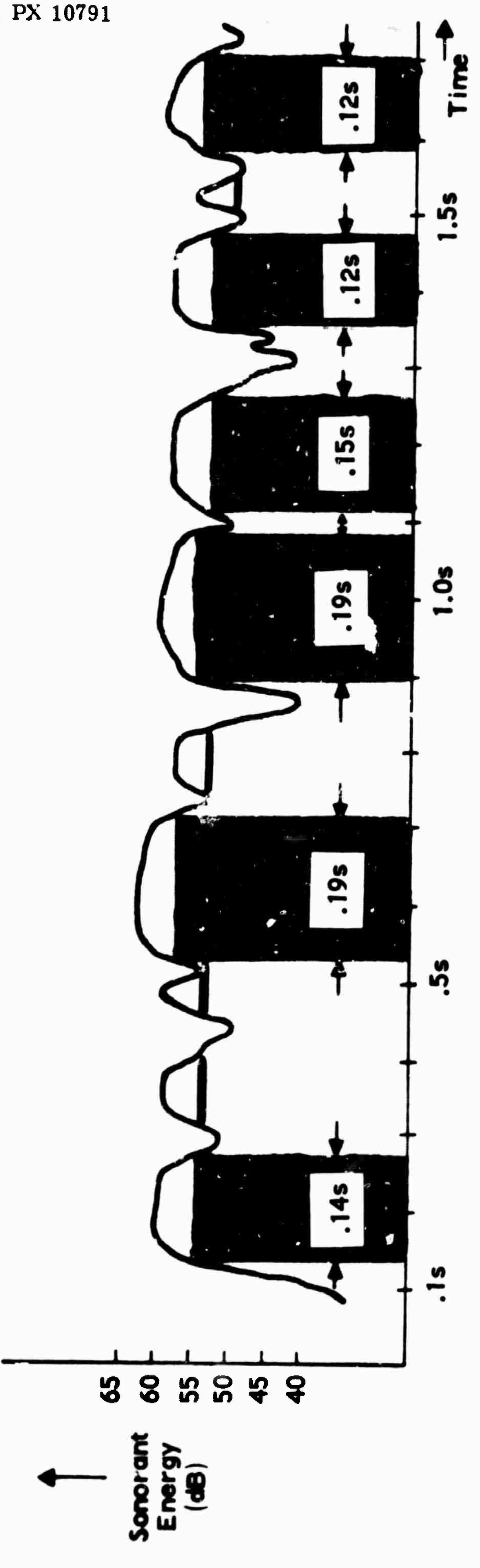


Figure 5. Location cf Stressed Syllables by High-Energy Chunks of Long Duration. Energy must remain within 5 dB of peak for .1 second.

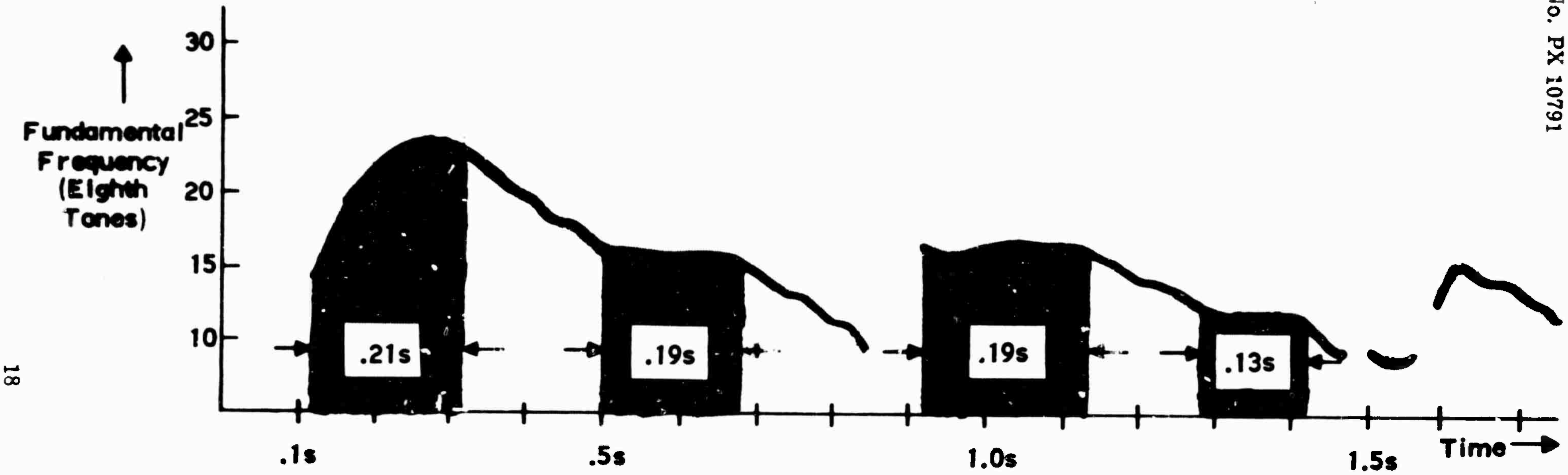


Figure 6. Location of Stressed Syllables by Regions of 'Increasing' Fundamental Frequency. Fundamental frequency must not decrease more than one eighth tone per .01 second. This must continue for at least .1 second.

TABLE III. PERCENTAGES OF STRESSED SYLLABLE LOCATIONS

(a) FROM DURATIONS OF HIGH-ENERGY "CHUNKS"

	RAINBOW (2 Talkers)	MONOSYLLABIC (2 Talkers)	ARPA (8 Talkers)
CORRECT	80%	94%	76%
FALSE	25%	25%	38%

(b) FROM INCREASES IN FUNDAMENTAL FREQUENCY

	RAINBOW	MONOSYLLABIC	ARPA
CORRECT	79%	84%	73%
FALSE	22%	23%	26%

(c) WITH ARCHETYPE-CONTOUR ALGORITHM

	RAINBOW	MONOSYLLABIC	ARPA
CORRECT	91%	93%	86%
FALSE	16%	22%	23%

tone per ten milliseconds (this is a relaxed form of a process of finding regions where fundamental frequency is steadily rising, or at least not falling rapidly).

Table IIIb shows that regions of increasing fundamental frequency are not as reliably related to stressed syllables as are the durations of high-energy "chunks", with poorest performance in the man-machine interactions of the ARPA Sentences. As shown by the corresponding results in Table IIIc, the archetype-contour algorithm obviously performs better than either of these two simpler algorithms, particularly for the spontaneous ARPA speech.

Table IV shows how stressed syllable location by the algorithms is affected by the type of sentence spoken (for the ARPA Sentences). For each algorithm, false alarms are most frequent in yes/no questions. The lowest correct location score from chunk durations occurs in yes/no questions, while the highest correct location score from increases in fundamental frequency occurs in yes/no questions. This suggests the value of combining the two types of cues to improve success in stressed syllable location, such as is done in the archetype-contour algorithm.

Figure 2 (page 8) shows how the archetype contour algorithm for stressed syllable location is being implemented within the acoustic analysis procedures at Sperry Univac. The beginning and ending of each stressed nucleus is determined, along with the positions of maximum energy within the nucleus and minimum energy between syllables (i.e., syllabic boundaries). Near the position of maximum energy (or "center") of the stressed vowel, the values of formants F_1 and F_2 are used, to categorize the stressed vowel as either high-front, low-front, low-back, or high-back.

Within the region surrounding the stressed vowel (bounded so as to include preceding or following consonants, but not to include surrounding syllables), analysis is done to locate all segments that might be an [r], an unvoiced stop, a nasal, a sibilant, or one of the "leftover" categories of silence, unvoiced consonant, or voiced consonant. This is actually accomplished using four independent phone detectors that look throughout the utterance for either sibilants, r-like sounds, unvoiced stops, or nasals. If in the stressed analysis region there are segments of sufficient duration that are not identified as being within the vowel, or within a sibilant, unvoiced stop, r-like sound, or nasal, those segments are categorized into the leftover categories of silence (very low energy), unvoiced consonant (unvoiced portion which is not a sibilant or stop), or voiced consonant (voiced portion which is not r-like, nasal, vowel, or voiced sibilant).

TABLE IV. EFFECTS OF SENTENCE TYPE ON STRESSED SYLLABLE LOCATIONS

		DECLARATIVES	COMMANDS	WH QUESTIONS	YES/NO QUESTIONS
ARCHETYPE ALGORITHM	Correct	88%	81%	87%	93%
	False	13%	23%	9%	30%
DURATIONS OF HIGH-ENERGY CHUNKS	Correct	79%	74%	83%	68%
	False	29%	39%	37%	49%
INCREASES IN FUNDAMENTAL FREQUENCY	Correct	72%	71%	70%	82%
	False	21%	23%	24%	38%

Thus, the entire time within the region of the stressed syllable is categorized into categories of phonetic segments. Only within the region of the reliably-encoded stressed syllable is such complete segmentation and classification attempted. Unstressed or reduced vowels, for example, are not even classified. In the next section, we shall see that only the most reliably encoded speech sounds are classified in the unstressed portions of speech.

2.2.3 Stress-Independent Phonetic Analysis – Not all of the reliably encoded information is in the tressed syllables. Some phone categories can be detected with fairly high and somewhat uniform reliability, regardless of where they occur in an utterance. Unvoiced sibilants were found to be correctly located in 66% of their occurrences in reduced syllables and 86% of their occurrences in unstressed syllables (Lea, Medress, and Skinner, 1973b), and have proven repeatedly to be among the phone types most reliably located in stress-independent phonetic recognition (Hughes and Hemdal, 1965; Medress, 1969). The characteristic low third formant of r-like sounds makes them among the easier sounds to recognize. The silences and sudden releases of unvoiced stops are also expected to be fairly easily detected. While sonorant consonants are among the most difficult phones to detect, and certainly difficult to distinguish among (cf. Newman, Li and Fu, 1973; Lea, 1973e), their recognition scores were found to be almost unaffected by stress levels (Lea, 1973e). In fact, the syllabic non-vowel sonorants ([l̩, m̩, n̩]) which occur in reduced syllables are among the easiest non-vowel sonorants to detect (Lea, 1973e). Nasals are expected to be easier to locate than [l]'s or glides [y, w].

We presently use only four independent phone-class detectors which search throughout the utterance for occurrences of each type of phone. Thus, sibilants [s, ʃ, tʃ, z, ʒ, dʒ] are found wherever high frequency frication energy (HE) is sufficiently greater than low frequency energy (LE) for a certain duration. The sibilants are further classified as alveolar or palatal on the basis of the frequency region wherein the highest energy concentration occurs. Also throughout the utterance, a search is made for releases of unvoiced stops, indicated by silence followed by a rapid spectral change (high value of a spectral derivative), with accompanying proper voice onset time to rule out glottal stops and voiced stops. The r-like phones are located wherever the third formant dips quite low. Nasals are indicated by a very low first formant of wide bandwidth. Then, all outputs from these detectors are intermeshed in time, to form an incomplete phonetic sequence which might be compared with expected sequences for words or phrases in the lexicon. As more sophisticated and more reliable procedures for phone detection and categorization are developed, these phonetic classification procedures will be augmented, but always with more weight given to the reliably-encoded stressed syllables and to those few phone categories which can be reliably and readily detected.

In the very simplified hypothetical version of a sentence lexicon and lexical matching process shown at the right side of Figure 2, all the phonetic segments found are first ordered in time, and a comparison is made with lexical entries that specify only those structural and phonetic aspects that the analysis procedures can provide. In addition to such phonetic information, the overall structural features of number of syllables, positions of stressed syllables in the spoken sentence, and the number of phrases or constituents might be used to determine the identity of a sentence.

To illustrate the use of the partial phonetic and prosodic specification and to illustrate the lexical matching procedures, let us consider a specific case of the sentence "Who is the owner of utterance eight?" The primary lexical entry for this sentence would assert that it consists of 3 constituents, with the sequence SOOSOOSOOS of stressed (S) syllables and other (O) syllables. The lexical specification of the phonetic sequence could be -[UC][HBV]x[SIBA] x [LBV][NAS][r] x [LBV][VC][r] x[NAS][SIBA] [LFV][STOP] -. Here the stressed vowels are identified as high front (HFV), low front (LFV) and low back (LBV), and the nasals (NAS), anterior [SIBA] and palatal [SIBP] sibilants, r-like elements, unvoiced stops, unvoiced and voiced consonants (UC and VC), and silences (-) are indicated, while segments that are not categorizable are indicated by x's. Any structure which met these conditions on specified segments would be acceptable. Also, additional lexical entries for the same sentence can be introduced, to allow for distortions like failures to detect a nasal, or vowel categorization errors. Such additional lexical entries might later be generated by rules.

Obviously, the lexical specification of sentences as such total units would be, at best, a temporary expediency which would be removed as techniques for constituent-by-constituent analysis, parsing, semantic analysis, and analysis-by-synthesis are developed. It should be apparent, however, from the acoustic analysis procedures shown in Figure 2, that prosodic features can provide independent structural information about the likely identity of a sentence, as well as guiding one to the most reliable phonetic information suitable for recognizing a sentence.

2.2.4 Rhythmic Aids to Phonological Analysis – Not explicitly shown in the general speech understanding strategy of Figure 1 or the acoustic analysis portions in Figure 2 is the possibility of using rhythmic information and measures of the rate of speech to guide phonological analyses. Two questions arise in considering the use of rhythm and rate of speech in speech understanding strategies: (1) How does one measure the rate of speech and detect rhythms? and (2) How can such information be used in speech understanding strategies?

It is reasonable to assert that useful acoustic measures of rhythm may be the time intervals between stressed syllables and the time intervals between constituent boundaries. Another possible acoustic measure of rate of speech is the number of syllables occurring per unit time, which will be available from our syllabication program. In Section 3.1 some results will be presented from Sperry Univac's initial studies of such acoustic measures of rate of speech and rhythm. Here we shall consider potential uses of such measures in speech understanding procedures.

Acoustic measures of rate of speech and rhythm may provide a potential check on the accuracy of stressed syllable location. If stressed syllables have been occurring at approximately equal intervals (of 0.4 seconds, say) and an interval of twice that expected interval (say 0.8 seconds) is found, a search for a stressed syllable that might have been missed could be initiated. If none is found, or if a pause is found to have intervened, the likelihood of certain structures may be thereby modified.

Rate of speech also has specific effects on other prosodic patterns, such as altering the amount of variation in F_0 during a sentence. A sentence spoken at high speed tends to have less variation in F_0 values (from maximum to minimum in the sentence) than its slow counterpart. This fact may be useful in on-line adjustment of thresholds in the constituent boundary detector or the stressed syllable location algorithm, so that less F_0 variation is demanded at higher speeds. (However, if such speed adjustments are used to alter boundary detections or stressed syllable locations, the original measure of rate of speech must either come from some independent process such as counting syllables per unit time, or else an iterative, feedback approach to boundary detection or stressed syllable location would be required.)

Also, the rate of speech may be used to modify expected relationships between prosodic and phonetic information, on the one hand, and underlying phonemic interpretations and word hypotheses, on the other hand. For example, at faster speaking rates certain slurring and coarticulation rules may apply, whereas more canonical unaltered phonetic sequences may be predicted to occur for slowly spoken utterances. An example would be the contrasts between "let's go eat" spoken at slow speed, " 'ts ko eat" at a faster speed, and "skweat" at even faster speed. The phonological rule component of a speech understanding strategy could be made to apply different analytical or generative rules to predict underlying sound structure, depending upon whether the speech is fast, moderate, or slow.

We plan to study whether rates of speech, as acoustically measured (cf. Section 3.1.), could be used to help improve either prosodic or phonological analyses. By measuring the time intervals between stressed syllables and between constituent boundaries, and the number of syllables per unit time we hope to provide potential cues to speech rate. These will be compared to phonetic segmentations provided by ARPA systems contracters, and both correlated with underlying phonological forms and necessary phonological rules, to see if speech rate could usefully guide selections of rules needed to arrive at underlying lexical forms.

2.3 Phonetic Aids to Prosodic Analysis

While the systems structures shown in Figures 1 and 2 show several ways in which prosodic features may be used to guide efficient and reliable phonetic analyses, it is worth noting that segmental phonetic information can in some ways also help prosodic analyses. For example, it is well known that the fundamental frequency contour and energy levels in a vowel, and the duration of the vowel, are all affected by the tongue height during vowel articulation (Lehiste, 1970; Lea, 1972, Chs. 4 and 5; Lea, 1973c). High vowels have higher F_0, lower energy, and shorter duration than low vowels, for physiological reasons. These phonetically-determined differences in such prosodic features are of substantial, perceivable size (10% or more), and may affect decisions such as whether or not a syllable is stressed. Thus, a high vowel which is more stressed than a low vowel could conceivably appear less stressed, since its duration and energy values are lower due to its intrinsic phonetic form.

Also, phonetic sequences have prominent influences on F_0 contours (Lea, 1972; 1973c), so that F_0 may fall in the initial portions of a stressed vowel following an unvoiced consonant, and the F_0 rise sought for in the stressed syllable nucleus may thus be missing. The stressed syllable location algorithm could thus fail to locate such a stressed syllable.

It is conceivable that adjustments could be made to compensate for such phonetic interference with prosodic patterns. For example, if formant tracking reliably shows that F_1 is low in a vowel, the vowel is high and the threshold values on energy and syllable nucleus duration could be lowered in the test for a stressed syllable there. An adjustment (in the proper direction) for energy, F_0, and duration thresholds in stressed syllable location could be made a direct function of the value of F_1.

In a similar fashion, knowledge of the presence of a voiced obstruent in a particular region of an utterance could be used to tell the constituent boundary detector that a local dip of F_0 in that region is not necessarily to be construed as a marker of a boundary between syntactic constituents.

In general, then, reliable knowledge of the phonetic structure of an utterance could be used to remove phonetic influences from prosodic patterns, leaving the readjusted pattern to reflect only (or at least primarily) the large-unit linguistic structure of the utterance.

2.4 Prosodic Aids to Word Matching

Prosodic features may be directly used in guiding procedures for locating and matching words in sentences. In long phonetic sequences, especially segment lattices (cf. Schwartz and Makhoul, 1974) which allow several alternative boundaries for segments, and which allow several segment labels to be assigned to various segments, it is difficult to tell where a candidate word might begin and where it might end. The location and phonetic specification of stressed syllables may provide reliable anchor points around which the search for an adequate word match might be attempted. Particular weight might thus be given to the most reliable information in the word. Also, words that might appear to be possible candidates for insertion at various points in the phonetic sequence may be ruled out if they have the wrong stress patterns. (To rely on this identification of lexical stress with stress as it appears in the context of a sentence, we must have reliable rules for relating lexical stress to sentence stress, such as Chomsky and Halle's (1968) rules are supposed to provide.)

We plan to investigate the use of located stressed syllables to aid the BBN word matcher (Rovner, Makhoul, Wolf, and Colarusso, 1974), which uses segmentation lattices and could be amenable to the use of stress in selecting candidate words which have stressed syllables in the correct positions. Thus, stressed syllables could be taken as "anchors" or islands of reliability around which the search for acceptable segment sequences in the segmentation lattice could be initiated. This could improve the efficiency of word matching from that currently being obtained with the "unanchored method", which looks everywhere in the lattice for matches of three-phoneme sequences that could correspond with lexical entries. Stressed syllable locations can also be effectively used in other procedures for word matching (Ritea, 1974; Weeks, 1974; Lesser, et. al., 1974).

2.5 Phonological and Prosodic Rules

Phonological rules form a major component in the general speech understanding strategy shown in Figure 1. They should characterize the alternative pronounciations possible, or likely, for underlying lexical representations in the context of continuous speech. A Sperry Univac representative participated in the ARPA Rules Workshop last January, and will be participating in future rules workshops. One contribution we hope to make is the compiling of published hypotheses and experimental confirmations about prosodic regularities, along with our own rules or hypotheses to be tested with the designed speech texts to be described in Section 3.1. There are a variety of published hypotheses about regular prosodic patterns in English, and about how such patterns are related to syntactic structures, semantic structures, phonetic sequences, and extralinguistic factors such as talker differences, emotions and physiological processes. Many of these hypotheses have not been tested with extensive speech data, yet they suggest many ways in which prosodic features might provide important structural cues suitable for aiding speech understanding systems.

One primary goal of our studies of published hypotheses and of new hypotheses is the specification of English intonation rules. We hope to determine all aspects of linguistic structure that can readily be determined from F_0 contours. In an earlier paper, Lea has outlined how F_0 contours generally seem to be composed of superimposed effects of clause structures, phrases, stress patterns, and phonetic influences (Lea, 1973c), but studies with the designed sentences will permit refining these gross regularities, and adding information about other influences on F_0 contours.

Word boundaries are an as-yet-unexplored aspect of structure that prosodic patterns may be useful for determining. The designed sentences (cf. Section 3.0) include minimal pairs for some word boundary studies, such as "in accuracy" versus "inaccuracy", or "an ice" versus "a nice", etc.

2.6 Prosodic Guidelines to Parsing

The number of syntactic boundaries in a sentence, and the numbers and positions of stressed syllables in the delimited phrases, may be very useful for selecting likely sentence structures and wording in phrases. An utterance with ten detected constituents would be highly unlikely to be an occurrence of a sentence structure with only three or four constituents. First selections in lexical matching could be done for words with stressed syllables in the correct places, and when a particular word (e.g., a noun) is confidently located, the approximate durations of preceding or following parts of a phrase may suggest whether or not one could expect other words (such as adjectives, articles, or second-portions of compound nouns) in the phrase.

It has repeatedly been asserted that English sentences have two basic intonation contours: Tune 1 for complete declaratives, commands, and WH-questions; and Tune 2 for Yes/No questions, and for uncertainty and incompletion (cf. e.g., Armstrong and Ward, 1926; Lieberman, 1967). Each type of contour is said to have a gradual rise in pitch to a peak near the first stressed syllable, followed by a steady fall from one stressed syllable to the next, with pitch in unstressed syllables either being somewhere between the values in surrounding stressed syllables, or else lower. The Tune 1 ends in a rapid fall in pitch after (or during) the last stressed syllable, while Tune 2 has a brief terminal rise in pitch in that same area. Alternative physiological explanations for these contours have been given (Lieberman, 1967; Ohala; 1970), but the important point here is that the sentence type (Yes/No question or not) is purported to be marked by the F_0 contour.

Thus, one could then have an acoustic cue to sentence type. This cue could, for example, rule out the possibility of the erroneous recognition of a Yes/No question as a command, such as has occurred for BBN's question, "Have any people done chemical analyses on this rock?" being incorrectly recognized as "Give any people done chemical analyses on this rock." (which is, incidently, ambiguous, increasing the gravity of the error). The terminal rise in F_0 expected in the yes/no question could conceivably be used to rule out the erroneous interpretation. (In Section 3.2, we will consider some actual evidence of the prosodic differences between these specific sentences.)

Prosodic features also provide cues to the boundaries of sentences in discourse, and to the number of clauses in each sentence (Lea, 1972; 1973b), which might be used to locate sentences and to rule out possible sentence structures.

Intonation contours have been asserted to provide data for "the recognition of immediate constituents and parts of speech syntax" (Trager and Smith, p. 77), such as distinguishing between "I'll move, on Saturday." and "I'll move on, Saturday.", or the classic "They [are] [flying planes]." versus "They[are flying][planes]." (Lieberman, 1967; Chomsky, 1957). Stress patterns or disjunctures are said to distinguish between "light housekeeper" and "lighthouse keeper", or between "large[tea cup rack]" and "[large tea cup]rack" (Lieberman, 1967, Chomsky and Halle, 1968). When two simple statements like "I saw a gray house" and "Joe saw a black house" are conjoined, to yield "I saw a gray house and Joe saw a black house.", the indicated shift of highest stress or emphasis occurs, showing a regular effect of coordination on stress patterns (Gleitman, 1965).

Published hypotheses and experimental results about so-called "comma intonation" patterns suggest that at the end of nonterminal clauses a brief rise in fundamental frequency will occur, essentially as an indication of imcompletion. The following two sentences have only one difference in vowel, and thus might be confused in a speech recognition system:

(1) After they met, John and Bill put the money on the table by the door.

(2) After they meet John and Bill, put the money on the table by the door.

The first sentence will usually be spoken with a terminal rise in the vicinity of the word met, followed by a possible pause, a large rise in F_0 at the clause – initial word John, and the word put unstressed or at least not highly stressed. Sentence (2), in contrast, will likely be spoken with a clause-marking terminal rise on the word Bill, a possible pause after Bill, a large rise in F_0 at the clause-initial word put, and with put stressed.

Other syntactic and semantic structures have accompanying distinctive prosodic patterns, such as the pauses, distinct "comma intonation" contours, and rhythmic effects produced by paranthetlcals or appositive (non-restrictive) relative clauses, such as in the sentence, "May, who knew the man in Rome, ran Maine." (to be contrasted to "Men who knew the man in Rome ran Maine.", which is without such interruptive markers).

One of the most common problems in disambiguating written sentences by parsing routines has been the occurrence of ambiguous sequences of noun phrase followed by a prepositional phrase, followed by another prepositional phrase. One example is shown in the alternative bracketings of sentence 1, shown in sentences 1(a) and 1(b):

(1a) After they met, John and Bill put [the money [on the table]] by the door.

(1b) After they met, John and Bill put [the money] [on the table [by the door]] .

The case where money which is already on the table is now to be moved near the door, may be contrasted to the case where money is to be placed on a table, which table is already near the door. One hypothesis worth examining is that a phrase which is subordinate under another phrase will have depressed values and reduced variation of fundamental frequency in comparison to its superordinate phrase. Thus, in one reading, the reduced relative phrase "on the table" will be more monotonically intoned, while in the other reading the reduced relative phrase "by the door" will be so monotonically intoned.

Stress patterns are also affected by placements of special emphasis and negation: for example, negation in an auxilliary verb will change its intonation contour, and the usual enclitic tie to preceding noun phrases can be lost (Lea, 1972; Allen, 1973).

With all these interacting aspects of syntactic and prosodic structure, and with the known effects of consonants and vowels on prosodic patterns (Lea, 1972; Lehiste, 1970), it is difficult to determine just what can be learned about sentence structures from prosodic features in arbitrary sentences. To isolate the various effects, one needs to compare prosodic patterns in utterances which are similar except for only one or a few separable differences. The test sentences (cf. Section 3.1) have been specifically designed to isolate effects due to sentence type, number of constituents, positions of stresses within constituents, various phrase categories, coordination, subordination, etc.

We plan to study subsets of the test sentences, to determine what aspects of syntactic structure can be determined even with little or no information about the phonetic structure of the words or "leaves" of a syntactic tree. These systematic experiments with the prosodic patterns in the designed sentences will be coupled with specific studies of how prosodically-determined syntactic structure can be used to aid syntactic parsers in speech understanding systems. For example, we hope to cooperate with BBN on developing ways of introducing useful prosodic data into their system (Woods, 1974), such as by using prosodic patterns to specify predicate functions attached to transition arcs used in parsing. Prosodic patterns may be used to assign priorities or likelihood information to various transition arcs, and to provide further qualifying information which must be satisfied before pop-up procedures are undergone. Similar introductions of prosodic information into other parsers could be attempted.

As each hypothesis about prosodic cues to syntactic structure is tested with the designed sentences and found to provide useful information, one can initially investigate its "a posteriori" use in disambiguating, or selecting among competing structures for an utterance. One may then also consider how that prosodic information can be introduced *early* in parsing procedures, to avoid fruitless paths of search in structural analysis, thus giving "a priori" guidelines to efficient parsing.

2.7 Prosodic Cues to Semantic Structures

There is some possibility of detecting aspects of semantic structure from prosodic patterns. It is known that emotion and some semantic distinctions (uncertainty, incompletion, doubt, etc.) affect intonation and other prosodies (Armstrong and Ward, 1926; Huttar, 1968). Some specially emphasized syllables are said to have accompanying *dips* in F_0 rather than the usual F_0 rises accompanying stress (Pike, 1945).

Also, grammatical relations such as coreference, contrast, antecedent-pronoun associations, etc., have been said by linguists to have regular effects on intonation. For example, Cantrall (1969) has suggested that there exists a form of "pitch concord" between noun phrases with coreference, so that in a sentence like (3), the pitch on "his" will be at different levels depending upon whether John, Bill, or Harry is the antecedent of "his":

4 2 6 6 4 2
(3) John told Bill that Harry had broken (his/his/his) bike.

Thus, if Bill's bike was broken, the pitch on "his" in the first sentence would be equivalent to that on "Bill", as shown in sentence (3) by Cantrall's pitch level 2 marked on both words. Likewise, consider sentence (4):

4 4 2 2
(4) Helen thought that when (she/she) went to Paris, Mary was pregnant.

we might determine whether Helen or Mary went to Paris by whether the "she" agreed in pitch with one or the other noun phrase. Cantrall has also claimed that so-called "alienable" possessions (e.g., his log") may be assigned a new pitch level in an English intonation contour, while "inalienable" parts of one's body (e.g., "his leg") effectively show at least an initial pitch concord with other references to the total person. This would yield such prosodic contrasts as those shown in sentences (5):

```
          4          4    6   3 4 4 5
(5)  Bill lost (his log/his leg) in the sawmill.
```

Finally, Cantrall has argued (concerning an issue that may be of major value in interpreting spoken utterances with gapping, agent deletion, etc) that deletions from grammatical structures are permitted in a sentence whenever the remaining parts can, by pitch concord, show who or what the intended referent is. Thus, the words "for her" or "for him" can be deleted, along with "herself" or "himself", in the sentence pair (6):

```
                                          2          2       2
(6)  John told Mary it might be fun { (for her)  to wash (herself) } in the fountain.
                                          4          4       4
                                    { (for him)  to wash (himself) }
```

because concord on the remaining verbal "to wash" helps identify the person being referred to. This has also been said to be true for deletion of passive "by phrases" which mark the agent of an action in a passive sentence.

Recent rules for stress assignment (Bresnan, 1971, 1972) assume that relative stress levels (determined by a "nuclear stress rule") are dictated by the embedded deep structures of sentences, which are applied through the iterative syntactic transformational cycle. Since deep structures are closely associated with semantic interpretations of sentences, stress levels (and thus their acoustic correlates) might then be relatable to underlying semantic structures. Some of these claims about semantic cues in prosodic patterns are to be instrumentally investigated, using subsets of the designed sentences.

There is a definite need to develop precise rules for systematically relating prosodic patterns to underlying structures. If one can understand how the interacting effects of semantics, syntax, lexical structures, stress patterns, and phonetic sequences are superimposed in the F_0 and energy contours of controlled English sentences, he has some of the most essential tools for using acoustic prosodic data to guide speech understanding strategies. As Woods (1974, p. 9) has noted, a prosodic component in a speech understanding system must know the "relationships between syntactic structure and meaning, on the one hand, and the intonation contour and stress patterns of a speech utterance, on the other." Woods continues by observing that, "When one considers the inherent ambiguity of the speech utterance which is entailed by the loss of word and phoneme boundaries and the relative uncertainty of identification of the elementary units of phonetic 'spelling', and when one contrasts this with the fact that sentences read aloud are capable of resolving syntactic ambiguities which are not resolvable in written form, it is clear that some additional information must be present in the spoken utterance beyond a mere sequence of vaguely blurred sounds. It appears that this additional information is provided in the subtle variations in pitch, energy, and segment duration which are present in the spoken utterance and which seemingly relate the speech signal directly to the syntactic structure of the utterance."

Only by a systematic attack on the task of compiling experimentally-verified rules can one hope to provide the kind of reliability needed to make such prosodic data of major value in speech understanding systems. The Sperry Univac study is now directed toward the initiation of such a systematic attack on the development and testing of useful prosodic rules.

3. EXPERIMENTS ON PROSODIC CUES TO LINGUISTIC STRUCTURE

3.1 Rhythm, Pauses, and Rate of Speech

3.1.1 Time Intervals Between Onsets of Stressed Vowels – English is said to be a "stress-timed" language, in that stressed syllables are purported to occur at approximately equal intervals of time. Allen (1972) has shown that listeners do in fact hear the "beat" of stressed syllables at *about* equal intervals, and that the beat occurs at the onset of the stressed vowel. At Sperry Univac, we have studied the distribution of sizes of *time intervals between the beginnings of the nuclei of stressed syllables*, for both perceived and algorithmically-located stressed syllables. Figure 7a shows the results for the perceived stressed syllables in the 31 ARPA test sentences. The number of occurrences of intervals of each duration is plotted versus the size of the time interval, with interval sizes quantized into fifty-millisecond increments. It is apparent that stressed syllables tend to be spaced about 400 milliseconds apart, although the variation in interval sizes is quite large. The wide variation might appear to be due in part to the fact that the 31 sentences involve eight different talkers, and a variety of speech styles from read speech to simulated man-machine dialog. The rate of speech could be affected by the task, the speaker, and the occurrence of hesitation pauses.

For more exact studies of time intervals between stressed nuclei, the effects due to talker differences, tasks, and other factors need to be carefully separated. The cross-hatched portion of the histogram of Figure 7a shows that portion of the total set of intervals that came from the speech of only one talker ("RB" designator on the ARPA sentences). Even with only one talker, the variation in intervals is considerable. Also, the Rainbow Script read by each of six talkers, and the Monosyllabic Script read by each of two talkers, provide sufficient data to determine whether the differences between talkers is a major source of variation in interstress intervals. Shown in Figures 7b to 7i are the histograms of interstress intervals, for each individual talker and text. While there is a gross clustering of intervals for these single-talker results, the variation in interval sizes seems at least as apparent for a single talker as for the pooling of eight talkers involved with the 31 ARPA sentences. This is also evident from the comparable values of standard deviation for the single-talker and multiple-talker results, as shown in Figure 7. We may conclude, with

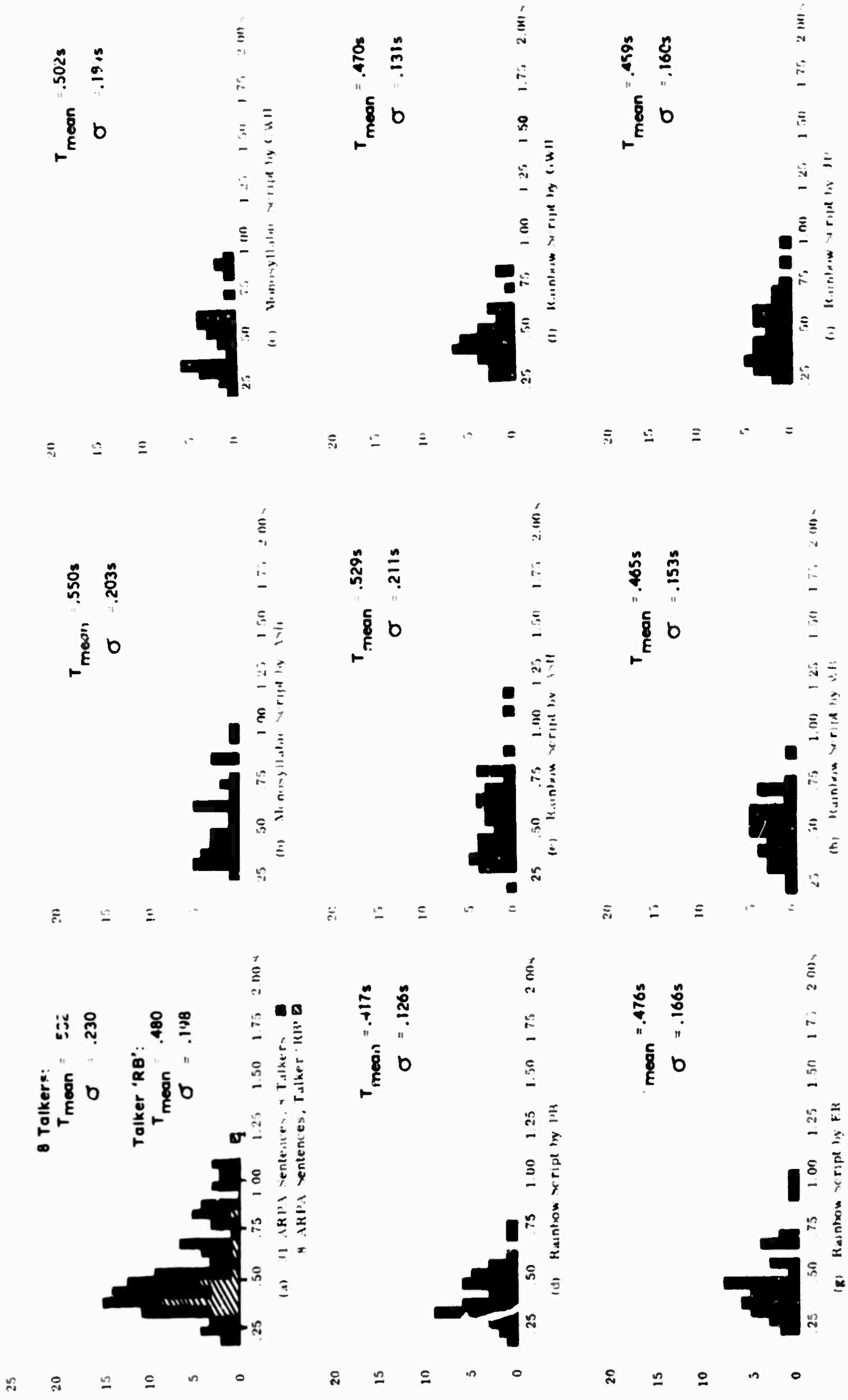


Figure 7. Histograms of the Number of Occurrences (Ordinates) of Various Sizes of Interstress Intervals (Abscissas, in Seconds). Mean Values (T_{mean}) and Standard Deviations (σ) are shown for each distribution.

Allen (1972, p. 4) and other researchers (Shen and Peterson, 1962; Bolinger, 1965; Abe, 1967; Allen, 1967) that the concept of English being stress-timed is not simply exhibited by exact equality of interstress intervals, or even by an unquestionable "tendency toward equality" of interstress intervals regardless of other factors. More precise hypotheses concerning the "tendency" toward isochrony need to be devised and tested. For example, some alternative hypotheses showing the role of unstressed[1] syllables in isochronism are the following:

1. Time intervals (Δt_S) between vowel onsets in stressed syllables tend to be equal, clustering, perhaps in a normal distribution, around a mean value, T_{mean}, (that is, $\Delta t_S \approx T_{mean}$), with the standard deviation of the distribution much smaller than the mean value.

2. Time intervals Δt_S between stressed syllables tend to follow a function of the form $\Delta t_S \approx T_{none} + T_1 + T_2 + \ldots + T_N$, where N is the number of intervening (unstressed or reduced) syllables between the stresses, and where $T_{none} > T_1 > T_2 \ldots > T_N$. Thus, the time interval between two stresses is minimum at T_{none} when no unstressed syllables occur between the two stresses. It is incremented by a smaller amount by each new syllable that is inserted. There is thus a resistance to the unlimited expansion of the interstress interval by intervening unstressed syllables (in that $T_i > T_{i+1}$), and some resistance to the cramming of more and more unstressed syllables into a fixed interstress interval (in that $T_i \neq 0$, for $1 \leq i \geq N$).

3. Time intervals Δt_S between stressed syllables tend to be a linear function of the number of intervening unstressed syllables: $\Delta t_S \approx T_{none} + NT_1$.

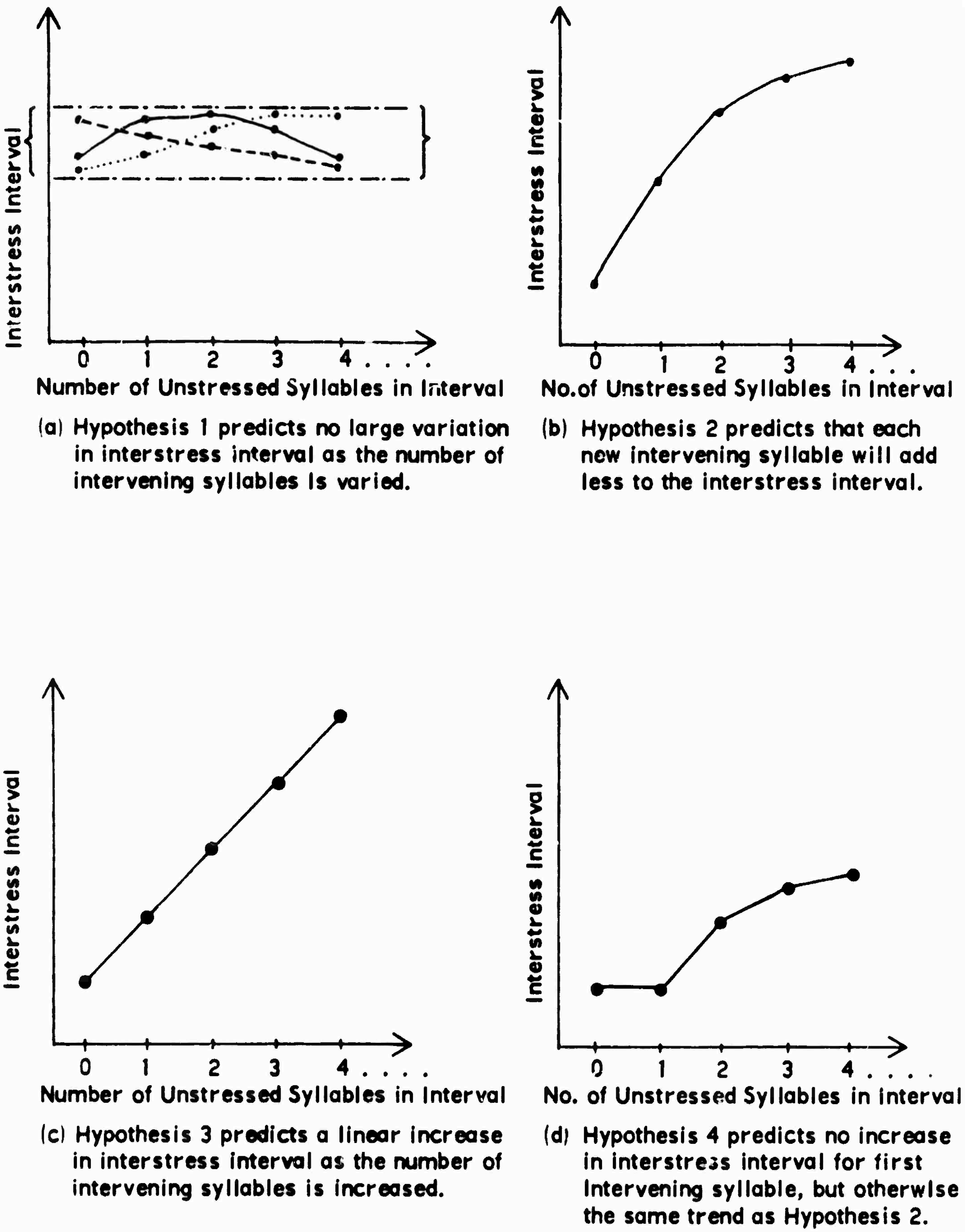
4. Time intervals Δt_S between stressed syllables tend to be the same with no intervening syllables or with one intervening syllable, but each additional intervening syllable after the first one adds a progressively smaller nonzero amount to the interval. Thus, in comparison to hypothesis 2, we have $\Delta t_S \approx T_{none} + T_2 + T_3 + \ldots + T_N$, where $T_{none} > T_2 > T_3 \ldots > T_N$, but $T_1 = 0$.

[1] In this discussion of intervening syllables between stressed syllables, the term "unstressed" will be used to refer to all syllables that were not perceived to be stressed (that is, they include syllables perceived as either unstressed or reduced).

Hypothesis 1 here is the simplest notion of a "tendency" toward equal interstress time intervals. It assumes that the number of unstressed syllables between two stresses has little or no effect on the interstress time intervals. The results shown in Figure 7, and previous results obtained by other researchers (Shen and Peterson, 1962; Bolinger, 1965; Abe, 1967; Allen, 1967), show that hypothesis 1 is, at best only marginally true. Bolinger (1965) has suggested a version of hypothesis 4 in which the "equality of interstress intervals containing zero and one unstressed syllable" is the source of our intuition that English is stress-timed. Allen (1968, p. 48) has also claimed that "an interstress interval with no unstressed syllables is just about as long as an interval with one stress in it", so that $T_1 \approx 0$. Allen (1968, p. 48) also proposed testing the hypothesis that the interstress "interval expands less and less as we add more and more syllables", as suggested by hypotheses 4 and 2. Pike (1945) described isochronism as the drawing out of unstressed syllables when there are few of them in an interstress interval and the jamming together of the unstressed syllables as their number increases between two stresses. This is the essence of hypotheses 2 and 4; namely $T_2 > T_3 \ldots > T_N$.

Hypothesis 3 asserts that the number of unstressed syllables already in an interstress interval has no effect on the duration of the increment introduced by an additional unstressed syllable. This disagrees with Pike's theory of jamming and drawing out, Bolinger's suggestions, and Allen's claims that it is common "knowledge" that an interstress interval is just as long whether it has one unstressed syllable or none in it.

Figure 8 shows pictorially the form of predictions made by these hypotheses. For each hypothesis, a plot is shown for the sizes of interstress intervals (ordinates) versus the number of unstressed syllables between stresses. Hypothesis 1 predicts that, regardless of the number of intervening unstressed syllables, the average interval size will be within some small band of values. Alternative curves connecting the average values for each number of intervening syllables are illustrated in Figure 8a. Hypothesis 2 predicts that each new syllable will add less to the size of the interstress interval, so the plot (Figure 8b) of average interstress intervals will be monotonically increasing, but it should level off as the number of intervening syllables increases. Hypothesis 4 predicts the same as hypothesis 2, but with no difference between zero and one intervening syllable, as illustrated in Figure 8d. Hypothesis 3 predicts a monotonically increasing plot with uniform slope (a straight line), as shown in Figure 8c.



(a) Hypothesis 1 predicts no large variation in interstress interval as the number of intervening syllables Is varied.

(b) Hypothesis 2 predicts that each new intervening syllable will add less to the interstress interval.

(c) Hypothesis 3 predicts a linear increase in interstress interval as the number of intervening syllables is increased.

(d) Hypothesis 4 predicts no increase in interstress interval for first Intervening syllable, but otherwise the same trend as Hypothesis 2.

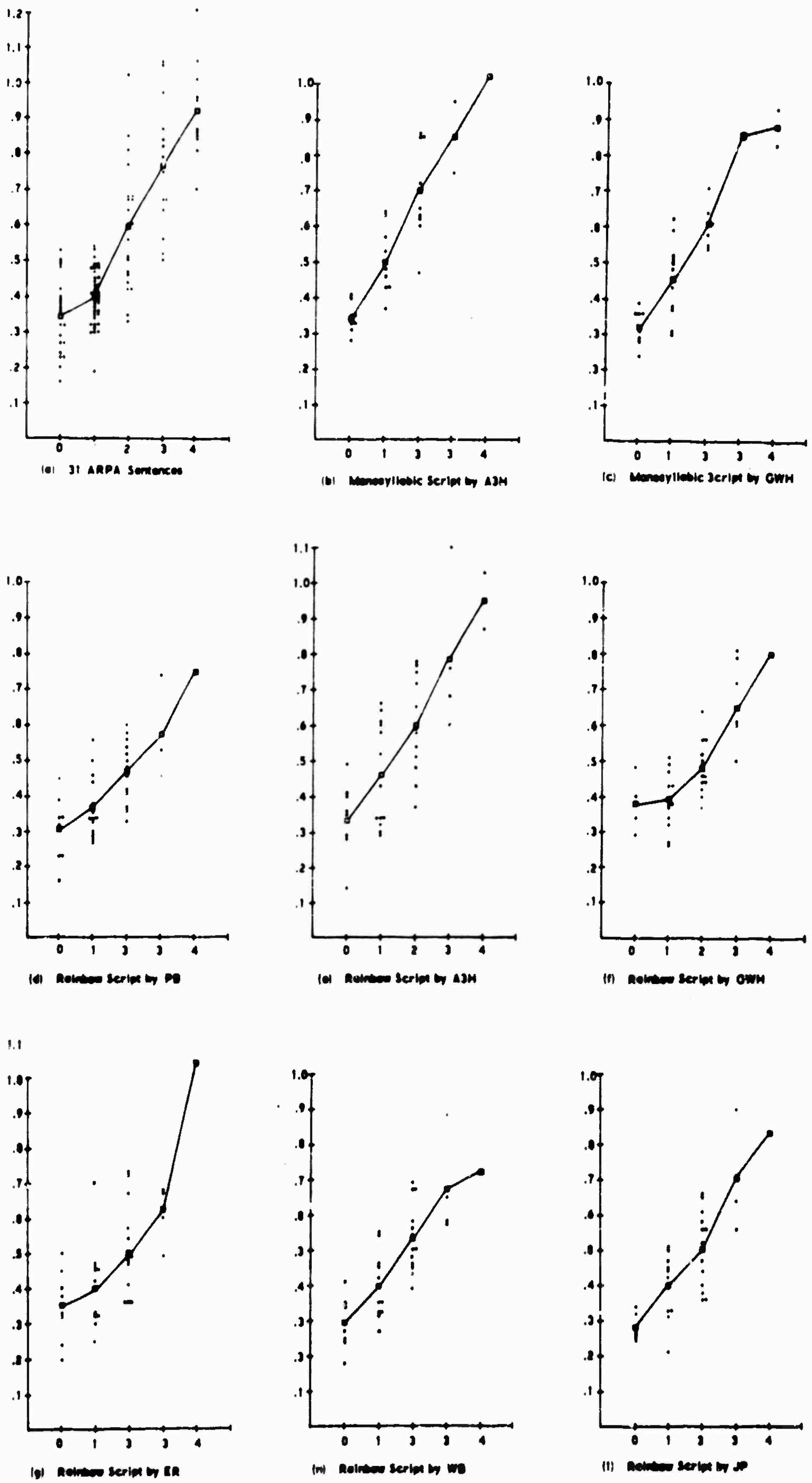
Figure 8. Plots of Hypothetical Relationships between Interstress Interval and the Number of Intervening Unstressed Syllables.

To evaluate these alternative hypotheses, the interstress time intervals in our speech texts were re-analyzed in terms of how many unstressed syllables were associated with each interval. Figure 9 shows, for each speech text, a plot of sizes of interstress intervals versus the number of unstressed syllables between the stresses. A dot is shown for each occurring interval, at the coordinates of its interval size and number of intervening unstressed syllables. (Thus, summations of all dots in the horizontal slice between, say, 300 and 340 ms would yield the height of the corresponding bar in the histograms of Figure 7.) The average value of interstress interval is computed for each number of intervening syllables (for each column in the graphs), yielding the values shown by the open squares. Lines connecting these average values show the trend in average size of interstress intervals as the number of intervening syllables is increased. Clearly, the average interstress interval increases substantially with each new unstressed syllable that is introduced between stresses.

The results clearly show that interstress intervals are substantially affected by the number of intervening unstressed syllables, in conflict with hypothesis 1. Hypotheses 2 and 4 are also apparently wrong for this set of spoken texts, since the plots of average interstress intervals do not level off with increases in number of intervening syllables. In conflict with hypothesis 4, there is an increase in average interval size as the first unstressed syllable is introduced.

On the other hand, the average values in Figure 8 do appear to agree fairly well with hypothesis 3, in that average time intervals appear to increase almost linearly with the number of intervening syllables, as predicted by the plot of Figure 8c. To be linear, the graph in Figure 9 should show uniform slope (i.e., equal changes in size of average interstress interval as the number of intervening syllables is successively incremented by one). In contrast, if the slope had decreased, this would have confirmed hypothesis 2 or 4, respectively. Since, in fact, the slope usually seems to increase with the number of intervening syllables, hypotheses 2 and 4 are clearly inadequate, and while hypothesis 3 is closer in performance, an even stronger effect of adding intervening syllables must be hypothesized.

While he never expected such results to ever occur, Allen suggested that any increase in the increment of interstress interval with each increase in the number of intervening syllables may be interpreted as showing "a resistance to the jamming



(a) 31 ARPA Sentences

(b) Monosyllabic Script by A3H

(c) Monosyllabic Script by GWH

(d) Rainbow Script by PB

(e) Rainbow Script by A3H

(f) Rainbow Script by GWH

(g) Rainbow Script by ER

(h) Rainbow Script by WB

(i) Rainbow Script by JP

Figure 9. Plots of Interstress Time Intervals (Ordinates, in Seconds) Versus the Number of Intervening Unstressed Syllables (Abscissas). [cf. the Hypothetical Relationships in Figure 8.]

together of unstressed syllables'' (Allen, 1968, p. 51). An alternative explanation is that the ideal stress pattern is one of alternating stress and unstress, as with all rhythms (Allen, 1968b; Miyake, 1902; Chomsky and Halle, 1968), and that as one inserts more and more unstressed syllables between two stresses, he tends to make one of the intervening syllables more like a stressed syllable, to re-establish something like the ideal alternation pattern. Thus, in going from one to two intervening unstressed syllables, an increment somewhat less than the total length of an unstressed syllable is introduced, because of the pull to keep it looking like a single intervening syllable. Then in going from two to three intervening syllables, a much larger increment occurs because of an attempt to make the middle syllable stressed (and thus quite long). Going from three to four intervening syllables reinforces the need for making a middle syllable stressed, because the pattern is so far from the ideal alternation; so the middle syllable might get lengthened some more, and the length of the fourth unstressed syllable would be added. Depending upon how much the middle syllable had already been lengthened by the insertion of the third syllable, and how much potential length was remaining to make that middle syllable look stressed, the size of the increment from three to four syllables may be more or less than the increment from two to three syllables. In Figure 9, we see both cases occurring.

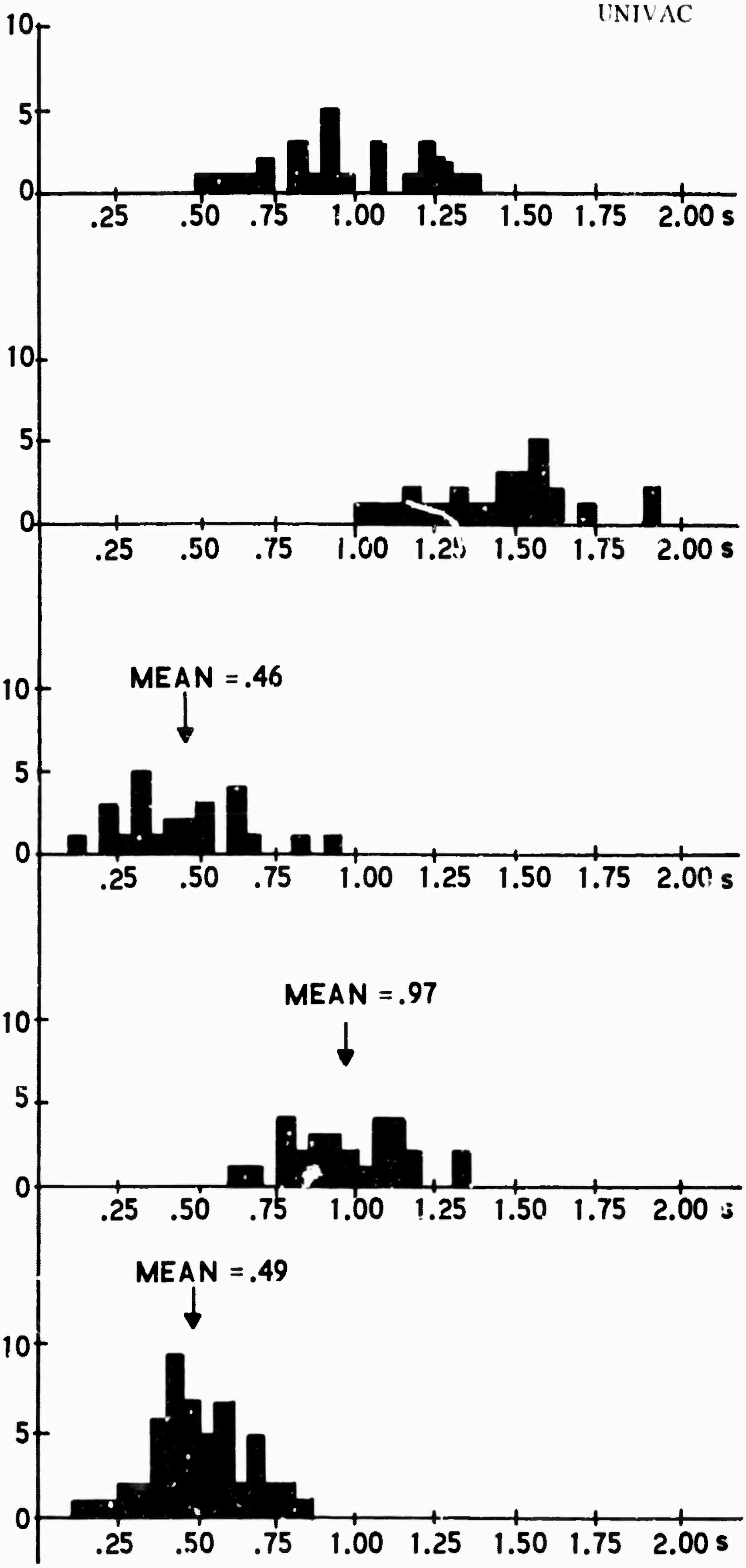
This hypothesis of a tendency toward stress-unstress alternation would suggest that whenever three or four syllables intervene between stresses, one of them (near the middle) would tend to look stressed. This was, in fact, the case for most of the occurrences of interstress intervals with three or four intervening syllables in the Rainbow and Monosyllabic Scripts. Of the 38 instances of interstress intervals with three or four intervening syllables, 21 included a syllable which was perceived as stressed by one (but only one) listener. Of the 17 remaining instances, 10 had syllables that were declared stressed by the stressed syllable location algorithm. These thus had some acoustic features, including sufficient duration, that made them appear stressed to the algorithm. The seven remaining syllable sequences were of the forms ing for the, -ently be-, -drops in the, and -bow is a di-. That is, they had no readily ''stressable'' syllables, without altering the normal stress patterns to an odd effect, such as putting a second stress within a lexical word, or stressing function words like prepositions or articles.

These results suggest the feasibility of a resistance to drastic variations from stressed-unstressed alternation, as do the plots of Figure 9. This agrees with other results that suggest that alternations are most prevalent in many human movements (Miyake, 1902; Woodrow, 1951). If stress is to be readily perceived, it must alternate with non-stress, so the contrast is most apparent.

3.1.2 Relationship Between Lengths of Pauses and Stress Timing – All of the results shown in Figure 7 and 9 are for stress intervals that do not span any pause in the speech texts. Though we have not explicitly noted it until this point, all interstress intervals spanning pauses have previously been excluded. This was done so the length of pause would not interfere with investigations of interstress interval in uninterrupted speech. Figures 10a and 10b show composite histograms, for all the Rainbow Scripts, of all interstress intervals that do span pauses. The pauses between sentences have been distinguished from those pauses at clause or phrase boundaries embedded within sentences. The histograms of pause durations are also shown in Figures 10c and 10d. ('Pause' here is defined as the duration of unvoicing in the speech, not silence, just as has been done in previous studies by Lea (1972).) Figure 10e is the result of subtracting the pause duration from each corresponding interstress interval, to yield a histogram very similar to those in Figure 7.

Notice the very close correspondence between the mean value (.47s) of interstress interval for the intervals not spanning pauses and that mean value (.49s) for those intervals spanning pauses after the pause duration is subtracted. We may conclude that pauses appear to be simple interruptions of the speech, with the remainder of the interstress interval of comparable size to that without the pause. The duration of the pause, on the other hand, is itself of interest. As shown in Figures 10c and 10d, the mean value of pause duration is very close to either the mean interstress interval, for embedded structural boundaries (.46s), or very close to twice that interval for boundaries between sentences (.97s).

3.1.3 Acoustic Measures of Rate of Speech – The mean time interval between stresses might be useful as a measure of speech rate. For example, the mean values in Figure 7 correctly suggest that talker ASH spoke more slowly than the other talkers, since his stresses were spaced farther apart. The mean values in Figure 7 did correlate with overall duration of the spoken text, so that those talkers who took longer to say the Rainbow Script showed the higher mean values of interstress interval. The problem



(a) Time Intervals Between Stresses When an Embedded Clause Boundary Is Included

(b) Time Intervals Between Stresses When a Sentence Boundary Is Included

(c) Durations of 'Pauses' at Embedded Clause Boundaries

(d) Durations of 'Pauses' Between Sentences

(e) Results of Subtracting Pause Durations from Corresponding Interstress Intervals that Span Those Pauses

Figure 10. Histograms of Interstress Intervals Spanning Pauses in the Rainbow Scripts Showing How Pauses Are Stress-Timed Interrupts in the Rhythmic Timings of Stresses.

of deciding whether interstress interval is a useful measure of speech rate is, however, complicated by the lack of any good standard for judging acoustic measures of speech rate. How does one decide for each new speech text, whether the talker is talking fast, medium, or slow? Also, how does one quantify the adequacy or accuracy of a specific acoustic measure of speech rate, and how does he objectively decide which of several alternative measures is 'best'? This remains an unresolved problem, which takes on some significance when we try to decide whether interstress intervals, number of syllables per second, or some other measure is best to use in speech understanding procedures.

It is worth noting that the boundaries between major syntactic constituents, as detected by the occurrence of fall-rise valleys in F_0 contours, also show considerable regularity in times of occurrence. Figure 11, for example, shows the histogram of time intervals between detected syntactic boundaries, for the 31 ARPA Sentences. Boundaries tend to occur at integral multiples of the mean interstress interval (about .4s and .8s). Occurrences of constituents which include one stressed syllable would presumably yield the first hump in the histogram, while constituents with two stresses yield the second hump, etc.

In addition to interstress intervals and intervals between constituent boundaries, another potential cue to rate of speech might be the mean intervals between onsets of any syllables (stressed, unstressed, or reduced). One would also expect that if English is really a stress-timed language and not syllable-timed (Pike, 1945), then the variation in sizes of intervals between onsets of all syllables would be substantially greater than that for stressed syllables (Allen, 1968). We are currently compiling results for onsets of all syllables in all our previously analyzed texts. This, of course, will provide information inversely equivalent to the measure of number of syllables per unit time.

These studies of rhythm and rate of speech will thus be continued and extended, particularly with the new designed speech texts. We must investigate criteria for selecting among acoustic measures of rhythm and rate, based on how such information will be used in speech understanding systems.

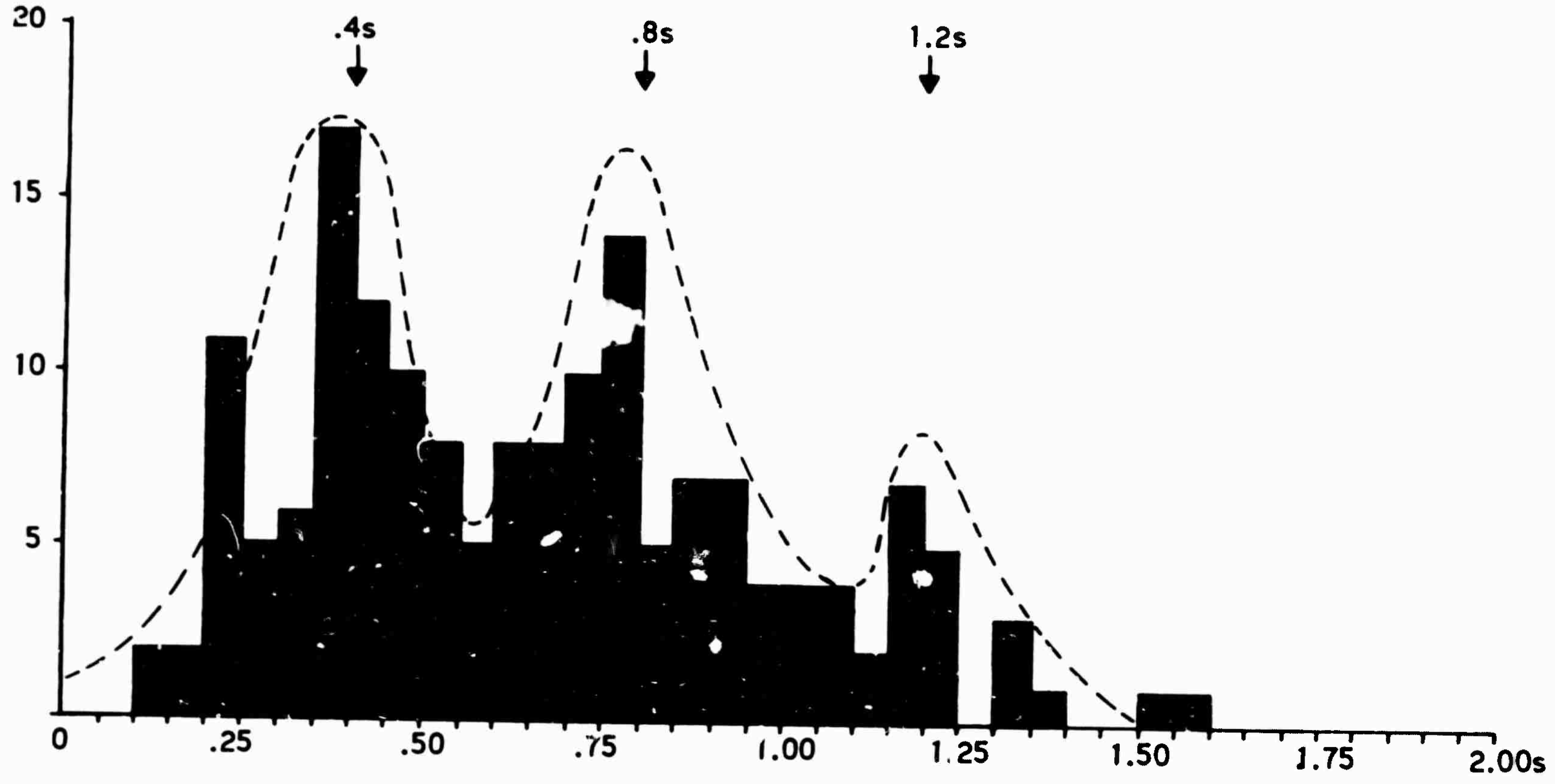


Figure 11. Histogram of the Number of Occurrences (Ordinate) of Various Time Intervals (Abscissas) Between Detected Constituent Boundaries, for the 31 ARPA Sentences. Boundaries appear to occur near Multiples of the mean interstress interval, as shown by the marks at .4, .8, and 1.2 seconds.

3.2 Preliminary Studies of BBN Problem Sentences

3.2.1 Sentences That Lead To Erroneous or Ambiguous Interpretations – In an actual demonstration of a speech understanding system, the Yes/No question (7) was erroneously recognized as beginning with the word "Give" rather than "Have", making it appear to be a command. Sentences (8a) and (8b) show that the wording of this command is ambiguous in structure:

(7) Have any people done chemical analyses on this rock?

(8a) Give[any people][done chemical analyses] on this rock.

(8b) Give any [[people – done] [chemical analyses]] on this rock.

Researchers at BBN have provided analog tapes of some such "problem sentences", for Sperry Univac to study with regard to the possibility of using prosodic cues to determine which structure was actually spoken.

Figures 12 to 17 show plots of the fundamental frequency and sonorant energy for each of six problem sentences provided by BBN. Figures 12 to 16,in particular, test the effects of various pronunciations related to sentences (7), (8a), and (8b). Figure 12 presents a naturally-occurring Yes/No question; Figure 13 presents another talker's pronunciation of the same sentence, but with declarative (falling) intonation; Figure 14 gives a version of sentence structure (8a) with a pause and prominent F_0 fall to accent the structural break; Figure 15 gives what appears to be an attempt for a "neutral" intonation, between structures (8a) and (8b); and Figure 16 gives an attempt at sentence structure (8b). Figure 17 gives an example of compound structures, to be discussed later.

3.2.2 Preliminary Results in Prosodic Disambiguation – It is apparent from the F_0 plot of Figure 12 that the terminal rise in fundamental frequency that is expected to accompany a yes/no question is not simply exhibited as increasing F_0 values within the last stressed syllable of the sentence. This is despite the fact that the sentence sounds like a question to the casual listener. If there is any cue in the F_0 contour to the distinction between a yes/no question (Figure 12) and other sentence types (Figures 13 to 16), it must be more subtle than a simple terminal rise in F_0. One other cue to sentence type that has been suggested (Ohala, 1970) is the general slope of the F_0

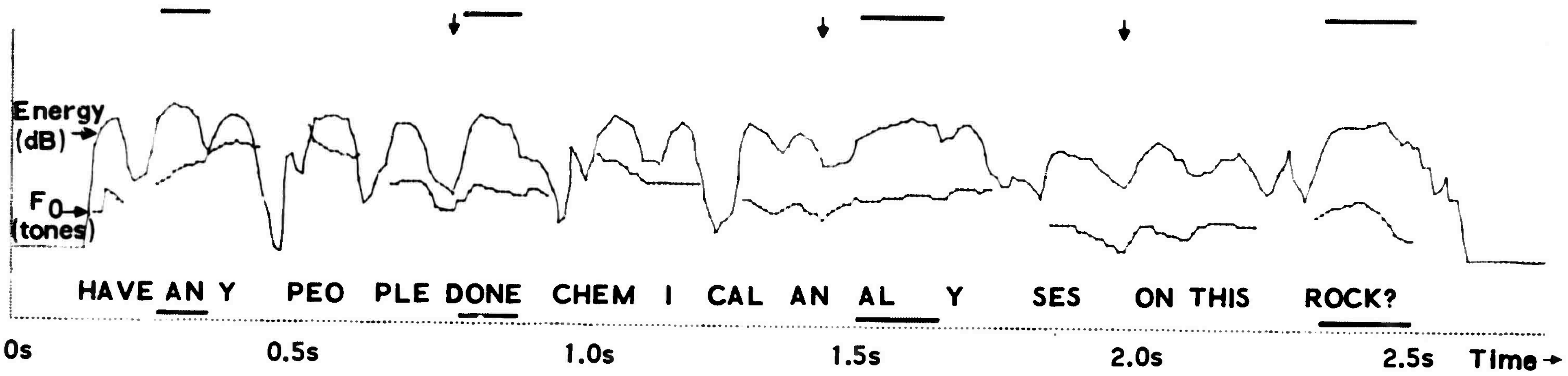


Figure 12. Plots of Fundamental Frequency (in Eighth Tones) and Sonorant Energy (in dB) for the Sentence DWD18A: "Have any people done chemical analyses on this rock?" Arrows mark detected syntactic boundaries, and horizontal bars mark the portions of the speech waveform declared (by the archetype contour algorithm) to be stressed syllables. Each syllable of the text of the sentence is shown approximately in the vicinity of its associated high-energy chunk of the energy contour.

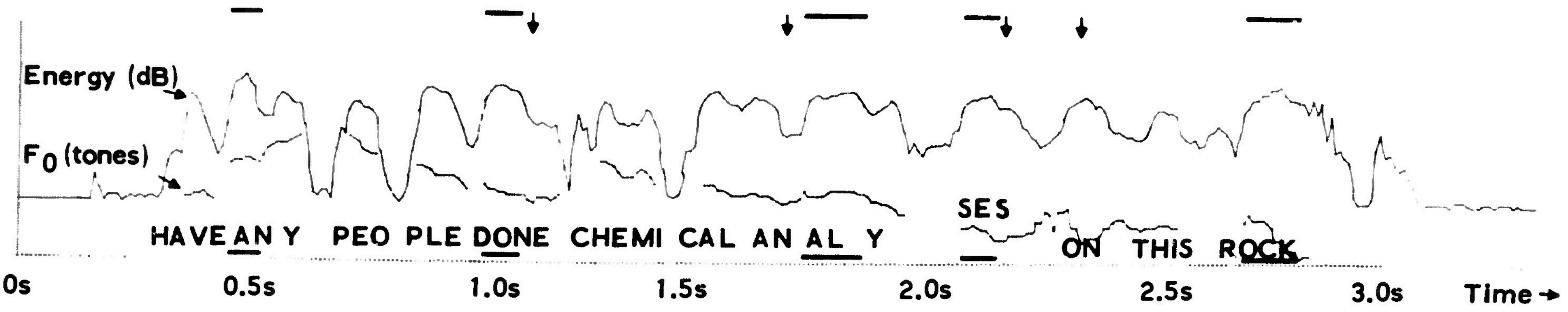


Figure 13. Plots of Fundamental Frequency and Sonorant Energy for the Sentence WAW18: "Have any people done chemical analyses on this rock."

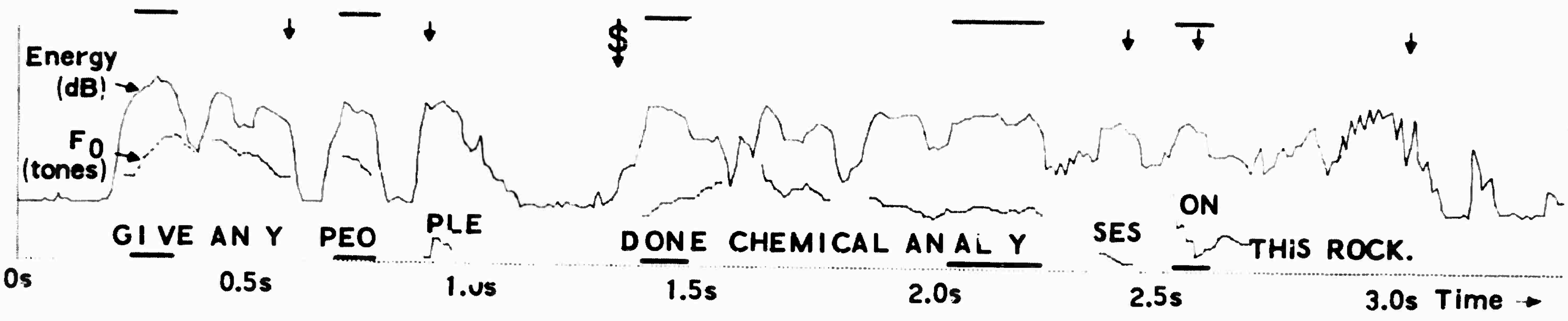


Figure 14. Plots of Fundamental Frequency and Sonorant Energy for the Sentence WAW20: "Give any people, done chemical analyses on this rock." ($ marks a sentential pause.)

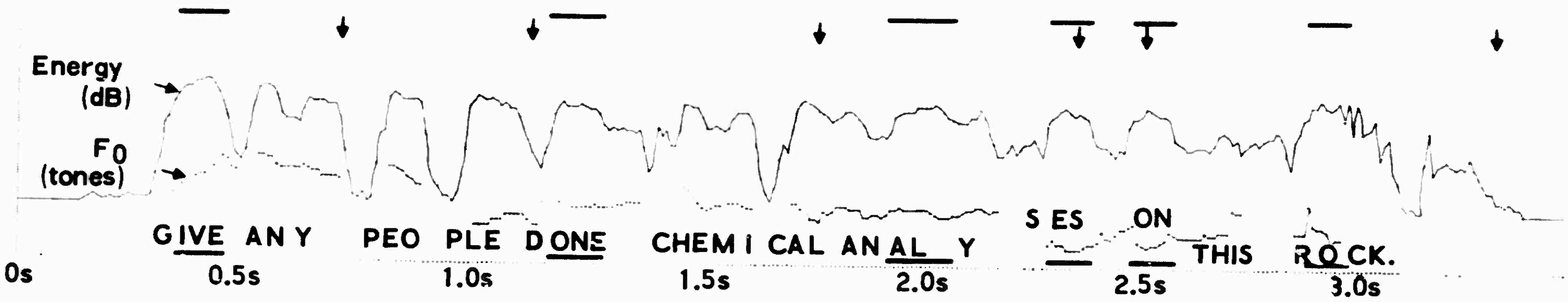


Figure 15. Plots of Fundamental Frequency and Sonorant Energy for the Sentence WAW21: "Give any people done chemical analyses on this rock."

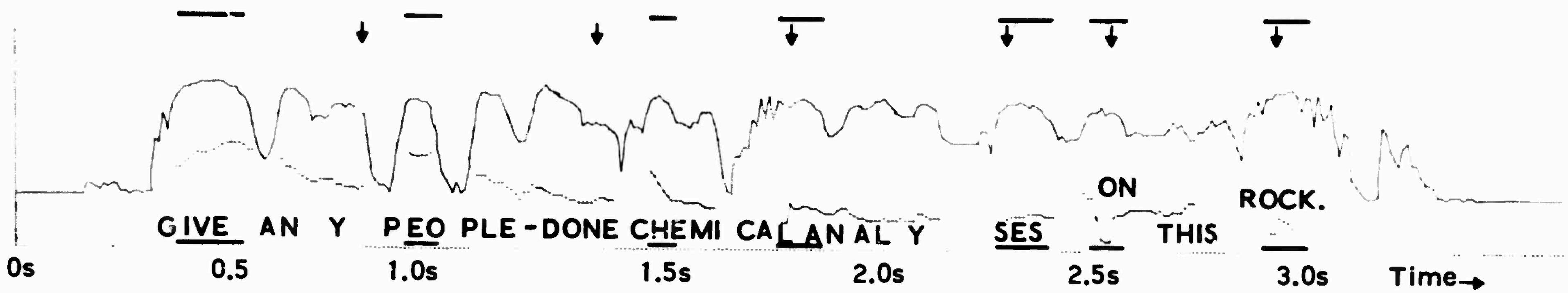


Figure 16. Plots of Fundamental Frequency and Sonorant Energy for the Sentence WAW22: "Give any people - done chemical analyses on this rock."

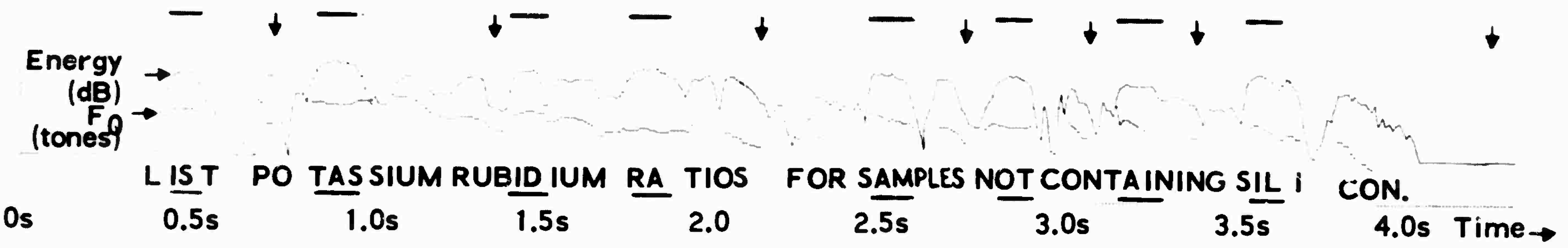


Figure 17. Plots of Fundamental Frequency and Sonorant Energy for the Sentence DWD25: "List potassium rubidium ratios for samples not containing silicon."

contour in the latter half of the sentence. It is noteworthy that the drop from the absolute maximum F_0 in each sentence to the maximum value in the last word ("rock") of the sentence is only 16 eighth tones for the yes/no questions of Figure 12, compared to 31, 48, 24, and 30 eighth tones for the sentences of Figures 13, 14, 15, and 16, respectively. An analysis of the 31 ARPA Sentences showed that the average drop in F_0 from the sentence maximum to the peak value in the last stressed syllable was 18 eighth tones for the commands, declaratives, and WH questions, and only 11 eighth tones in the yes/no questions. Thus, the degree of the gradual fall in F_0 after the first stressed syllable of a sentence might be a possible cue to sentence type. Obviously, further studies with far more data are required to adequately test such cues. The designed texts (cf, Section 3.3) will include extensive tests of effects of sentence type on F_0 contours.

Shown on the plots of Figures 12 to 17 are horizontal bars during the portions of speech that were declared to be stressed by a hand analysis with the archetype contour algorithm for stressed syllable location. Due to some local quirks in the F_0 contours, and consequent errors in constituent boundary detection, some false locations were found, such as for the unstressed syllables "-ses" and "on" in the word-sequence "analyses on this rock". Also, the syllable "chem-" in "chemical analyses" was not usually located as stressed, though one might expect that it would be perceived as stressed. The F_0 values in the vocal fry of some instances of the terminal word "rock" were too low to be determined by the F_0 tracker, and thus that stressed syllable was not located by the stressed syllable location algorithm. It may be appropriate to adjust F_0 tracking and stressed syllable location procedures to allow for the rapid F_0 fall in such terminal (pre-pausal) positions.

The stress patterns in Figure 12 to 16 do provide some cues to the intended syntactic structures. In particular, the auxiliary verb "have" in Figures 12 and 13 is unstressed, while the command verb "give" is stressed in all its instances in Figures 14 to 16. This is a useful cue as to whether the sentence is a command or a yes/no question, and may be especially useful if terminal rises in F_0 or other cues fail to reliably establish sentence types. We expect this contrast in stressedness of the first verbal to always be exhibited, except when special emphasis or contrastive stress introduces a stressed auxiliary in the yes/no question. Figures 12 to 16 also show that only in the wording of the yes/no question is the first syllable in "any" stressed.

Other stress contrasts shown in Figures 12 to 16 are structurally important. The word "done" might be expected to be stressed when it is the main verb of the yes/no question or the head word of the noun phrase "done chemical analyses", but not when it is the latter part of the compound construction "people-done" (ala Figure 16). Thus, the lack of stress on "done" could distinguish the construction of sentence (8b) from the other structures. The stress on "peo-" of "people" is most evident in the compound construction of Figure 16, as had been predicted (Lea, Medress, and Skinner, 1974, text of oral paper).

The positions of detected syntactic boundaries also should show interesting structural contrasts. The compound modifier "people-done" of Figure 16 is shown to be a structural unit by the boundaries surrounding but not internal to it. The only other case where no boundary occurs between "people" and "done" is in Figure 13, where there is no boundary between the noun phrase "any people" and the word "done". Past experience has shown that such NP-verbal boundaries are frequently not shown in F_0 contours. This is one of many forms of enclitic ties between noun phrase subjects and their verbs. The boundaries found between "people" and "done" in the sentences of Figures 12, 14, and 15 can be used to rule out structure (8b) as a likely interpretation for those sentences. The sentential pause (↓) marked in Figure 14 is a particularly prominent cue to structure (8a) versus (8b).

Another cue to structural boundaries is "disjuncture", defined by the time intervals between onsets of syllabic nuclei (Lieberman, 1967). Several differences in time intervals between syllabic nuclei in Figures 12 to 16 do show expected contrasts. For example, this time interval between "have" and "any" is 100 or 110 ms in Figures 12 and 13, compared to 160 or 170 ms for the stressed "give" in Figures 14 and 15, and 260 ms for the "give" (in the underlying sense of "Give to someone the following . . .") of Figure 16. This is a reflection of the unstressed form of "have" and stressed form of "give", and also of the distinctive intent of the "give" in structure 8b. Similarly, the disjuncture between "people" and "done" in Figures 14 and 15 is shown by 190 or 180 ms time intervals, compared to 140 or 150 ms for the other structures (Figures 12, 13, and 16) without such a disjuncture. The prominent stress on "chem-" in Figure 16 makes it be followed by a 320 ms time interval, compared to 230 to 250 ms for the same place in the other sentences.

In summary, then, there are several cues in overall F_0 contours, positions of constituent boundaries, stress patterns, and disjunctures that reflect significant structural contrasts such as sentence type and structural bracketing. These show promise of distinguishing yes/no questions from commands and of disambiguating structurally ambiguous sentences. Each such cue did suggest, for example, that the structurally ambiguous (or intended "neutral") pronounciation of Figure 15 looks most like structure 8a, and not 8b. The yes/no question in Figure 12 might be guessed to be a question (thus avoiding the former confusions found with such a sentence), if the stress of the first word, its following lack of disjuncture, and the overall F_0 contour are given consideration in selecting the first hypotheses to try in the parser.

Figure 17 shows another sentence of interest to BBN. As a command, its verb is stressed, and in fact the located stressed syllables are exactly those that might have been expected to be stressed. These include the negative particle "not". The compound noun "potassium rubidium ratios" is broken into two detected constituents, which has previously been shown to occur for some, but not all, compound constructions. Further studies of F_0 contours in compound constructions are needed. Figure 17 also includes a relative structure of the multiply embedded form [NP [PP[VP]]. Such constructions warrent further study. For example, it is not immediately obvious from the F_0 contour that "for samples" is a relative under "potassium rubidium ratios", or that "not containing silicon" similarly modifies "samples".

3.2.3 The Need for More Controlled Studies – These studies of the six BBN sentences are only intended to illustrate the types of information that might be provided by prosodic patterns. No strong claims can be made from these preliminary studies, but the results with Figures 12 to 16 do reinforce our opinions that some aspects of linguistic structure can be usefully detected from prosodic features, and possibly used to guide syntactic hypothesizing. Further studies with extensive sets of utterances with controlled contrasts must be undertaken, to determine exactly what can be readily determined from prosodic patterns. As each prosodic cue to linguistic structure is firmly established from such experimental studies, that cue may be applied to actual tests with available ARPA speech understanding systems.

3.3 Design of Speech Texts

There is a definite need to develop precise rules for systematically relating prosodic patterns to underlying linguistic structures. Many hypotheses about regular prosodic patterns have been published, but few have been tested with extensive speech data. Studies of "near-minimal-pairs" of sentences, which are identical except for one difference in linguistic structure (or a few isolatable differences in structure), will help determine the correct form of the rules relating prosodic patterns to linguistic structures. A complete report on the design of sentences for testing prosodic, phonetic, syntactic, and semantic contrasts is in preparation. In this section, we shall briefly consider some examples of the designed sentences and their uses.

3.3.1 Sentences for Testing Phonetic Sequences – Sentences which include a comprehensive variety of phonetic sequences can be used to test phonetic effects on prosodies, to test the accuracy of phonetic recognition procedures, and to test parameterization procedures such as formant tracking. It particularly seems desirable to provide instances of all word-initial (#CV) and word-final (VC#) sequences, for stressed vowels.

Table IV shows some example sentences that have been compiled for including such #VC- and -VC# sequences. Eleven vowels have been included: /i, I, ɛ, e, æ, ʌ, ɝ, a/ɔ, o, ʊ, u/. Vowels /a/ and /ɔ/ are too readily confused by speakers of many American English dialects for us to demand their distinction in the texts. In most cases, each sentence tests a single consonant in either initial or final position, with five or six of the eleven vowels. Whenever it could be done readily, the stressed syllables were placed in monosyllabic words. A few #CV and VC# sequences don't occur in English words.

The arrangement of such similar phonetic structures in close proximity may cause "tongue twisting" and a strong sense of alliteration in the talker. To reduce such effects, and the possibility of somewhat atypical pronunciations, most pairs of words with similar structures have been separated by one or more words of a quite different phonetic structure. It is apparent that some degree of poetic, atypical pronunciation still remains. We have considered rearranging the words in the sentences so that in each sentence a few word-initial CV sequences are included with a few word-final VC sequences, for the same C. The phonetic similarities may be less obvious in such constructions.

TABLE IV. SENTENCES FOR TESTING INITIAL AND FINAL CONSONANTS (UNVOICED SIBILANTS AND STOPS)

Phone Category Being Tested	Sentences
#s——	Sue was sick of setting tables and serving salad and suppers. See if you can soak the soot from Sadie's sox.
#ʃ——	In old shoes and shirt, Sheila shopped for shawls at The Shack. In a close shave, the ship shook and shuddered from the shell, but showed no damage.
——s#	My boss made a fuss about this place having class and being close.
or	Puss is loose and making a mess of the fresh fish and hash.
——ʃ#	Which is worse to the peace of a pet: a wash, a brush, or a leash that won't reach the bush?
#p——	Pete won't pick a popular pet like a pup or pony. It doesn't pay for a person to panic or push near the pool.
——p#	That big ape can slurp up a cup of soup in a snap. We will sweep and mop the step, then wipe up the soap so you don't slip.
#t——	Ted took two tough courses, and taught each term. Toby typically takes tap water to make tea.
——t#	Your mutt sat on my suit with a lot of dirt on his feet. I'll bet they thought you put your vote in a bit late.
#k——	A clean kettle is the key to cooking a good cup of coffee. Kay can't quite cope with Kirk's parking the cab at the curb to kiss and coo.
——k#	To pick that lock will take work, a real knack, and a lot of luck. Did you check whether or not Duke broke the hook off the plaque?

The importance of having several instances of the same consonant in a single sentence cannot be overemphasized, when one considers the uses of these sentences. To test success in automatic phone categorization, for example, one would like to have a few sentences which include a number of instances of a particular phone, in various phonetic contexts. For ease of processing, for efficiency, and to maximize the number of questions answered by the fewest number of sentences, one would like to pack as many occurrences of relevant phonetic sequences into as few sentences as possible. On the other hand, each sentence should be of manageable duration, not inordinately long or complex, and not awkward to say. Making about every other word in a sentence include a relevant phonetic structure seems to be a reasonable compromise between efficiency and naturalness of expression.

In addition to testing a phonetic classification scheme with sufficient occurrences of the specific phones to be detected, it is useful to have a set of sentences which readily provide possible "false alarms" resulting from similar phones. Thus, to test a program for sibilant location, one could apply it to the sibilant sentences in Table IV and to sentences which contain other fricatives or which contain unvoiced stops whose releases may be mistakenly called occurrences of sibilants.

Sentences which test certain word boundary effects are also included in the designed set of sentences, though most of them are sentences which are also designed to study syntactic contrasts. For example, Table V lists some sentences which test for vowel hiatus effects. (Hiatus has been defined as "the occurrence of two vowel sounds without pause or intervening consonantal sound" (Webster's Seventh Collegiate Dictionary).) The difficulty of saying certain vowel-vowel sequences has led to the inclusion of English "anti-hiatic mechanisms" such as the inserted n in "an owl", word-initial glottal stops before vowels, and the occurrence of word-initial "h's" in many English words. Anti-hiatic mechanisms are also behind the alternation between [ðə] and [ði] for "the", so that "Who is the owner of utterance eight?" has [ði] rather than [ðə]. The occurrences of vowel-vowel sequences such as those in Table V provide means for studying these phenomena and writing relevant phonological rules.

Other word boundary effects to be studied include cues as to where word boundaries should be placed (e.g., in "a nice" versus "an ice", "inaccuracy" versus in accuracy", etc.)

TABLE V. SENTENCES EXHIBITING VOWEL-VOWEL SEQUENCES AT WORD BOUNDARIES

Underlined portions show vowel-vowel sequences, with the associated categories of both vowels. (High or Low, Front or Back, or ɝ-Like)

Mary Marie in May.
HF/HF

Marry Marie early.
HF/ɝ

Marry her early.
ɝ/ɝ

I am one and you are one.
HF/LF HB/LB

Really enroll in May.
HF/LF

The law about any umlaut has really allowed few errors.
LB/LB HF/HB HF/LB HF/LF

You are one.
HB/LB

You are in one.
HB/LB ɝ/HF

You lure Ann and Ron will Lure Amy into the room.
ɝ/LF ɝ/LF HF/HF

Is her umlauting in error?
ɝ/HB

Marry her anew.
ɝ/LB

We view umlauting as important.
HB/HB

Allow error.
HB/LF

Ron will enroll Lou early.
HB/ɝ

The law Ann learned was the law umlauting this vowel.
LB/LF LB/HB

Do Ann and Lou unnerve you and make you ill?
HB/LF HB/LB HB/LF HB/HF

Law easily confuses Ron.
LB/HF

Ron learned law early.
LB/ɝ

The designed sentences also include sentences with identical syntactic strucure but contrasting phonetic structure. Some sentences have only sonorant sounds in them, while others are designed to have only fricatives and vowels, or only stops and vowels, etc. Thus, one can compare triads like "Loan Ron rum.", Serve Sue fish.", and "Take Kay pop." One can also investigate interacting phonetic and syntactic contrasts such as in the declarative "Sue saw Fay." versus the WH-question "Who saw Fay?"

A forthcoming report will describe these various phonetic studies that can be done with the designed sentences.

3.3.2 Sentences for Controlled Studies of Prosodic Cues to Linguistic Structures. The designed sentences permit testing a variety of syntactic and semantic effects on prosodic patterns. For example, as mentioned in Section 2, studies are to be done on prosodic cues to sentence types with such groups of sentences as (9) to (12):

(9) Pete can take Kay.
(10) Can Pete take Kay?
(11) Take Kay.
(12) Who can take Kay?

Other syntactic questions to be investigated concern the effects of coordination of sentences, verb phrases, adverbs, and noun phrases, as shown in sentences (13) to (21):

(13) Ann and Ron rule Maine.
(14) Ron and Ann rule Maine.
(15) Ron knew Ann and May.
(16) Ron knew Ann, May, and Lou.
(17) Ron knew Ann and May knew Lou.
(18) Ron and Ann knew May and Lou.
(19) Ron and Ann knew May and Lou, respectively.
(20) Wayne, Ron and Ann knew Lynn, May, and Lou, respectively.
(21) Ham and eggs and pizza and beer are my two favorite meals.

Subordination of one phrase under another is also to be studied, with contrasting sentences such as (1a) and (1b) given earlier (cf. page 30).

The syntactic bracketing contrasts that were illustrated by Chomsky's "flying planes" paradigm are to be studied with unambiguous but contrastive sentences like (22) to (25):

(22) Lawmen are lying men.
(23) Lawmen are ruling Maine.
(24) Airmen are erring men.
(25) Women are airing wool.

The distinction between restrictive and non-restrictive (apositive) relative clauses, and the expected differences in F_0 contours, pauses, and rhythms, are to be studied with sentences like (26) to 31).

(26) Ann, who knew Ron, ran Maine.

(27) Men who knew Ron ran Maine.

(28) Ann knew May, whom Ron knew.

(29) Ann knew men whom Ron knew.

(30) "Will men," Ron moaned, "run Maine?"

(31) Will men Ron named run Maine?

Other sentences test effects of: adverbs; prenominal and predicate adjectives; possessives and quantifiers; pronouns; boundaries between noun phrases and verbals; negation; there-insertion; etc. Studies of stress patterns are to be extended to include noun-verb pairs, as shown in sentences (32) to (39), where underlining shows stress contrasts:

(32) Our new object increases inaccuracy.

(33) Verv few object to increases in accuracy.

(34) The two records permit a conflict in schedule.

(35) His records conflict with his permit in several ways.

(36) Let's record our progress to the tower.

(37) Let's progress towards record altitude.

The influence of the "tonalization position", or utterance final position in an intonation contour, is to be studied with sentences such as (38) and (39), where "mine" in terminal position is contrasted with "mine" in non-terminal position.

(38) Run mine.

(39) Run mine now.

The first issue to be systematically addressed in the planned study of the designed texts concerns the effects of the position of the first stressed syllable within a constituent. As stress is moved within a constituent, how do the associated positions of syntactic boundaries move, and how are the acoustic correlates of stress affected? Such questions are soon to be studied, with sentences like (40) to (49), where again underlining indicates stress:

(40) Marry me.

(41) Enroll me.

(42) Marry May.

(43) Enroll May.

(44) Marry Murray

(45) Enroll Murray.

(46) Marry Marie. (47) Enroll Marie.

(48) Marry Leonora. (49) Enroll Leonora.

All these designed sentences will thus permit careful study of the prosodic effects of various isolated differences in linguistic structure, and will provide the necessary information for development of experimentally-verified rules of prosodic structure.

3.3.3 Prosodic Analysis of Man-Machine Dialogues. The extensive study of the prosodic patterns in the designed texts will provide valuable information about the effects of semantic, syntactic, and phonetic structure on prosodic patterns, in carefully controlled contexts, so that the exact origin of any prosodic patterns can be determined. Yet, the reading of such designed texts by no means represents the exact form of speech expected in actual speech understanding systems. Man-computer interaction with speech is expected to be significantly different from read speech. It also should be different from casual conversations.

Consequently, we plan to analyze selected portions of simulated man-machine protocols, along with the more controlled studies of the designed sentences. Comparing prosodic patterns in read speech and man-machine speech should provide us with guidelines as to which aspects of the results from the controlled tests with the designed sentences can be directly translated into man-machine contexts, and in what ways other aspects must be altered to apply to man-machine dialogs. The necessary dialogs are to be obtained from other ARPA contractors who are recording such speech for use in their speech understanding systems.

4. CONCLUSIONS AND FURTHER STUDIES

Our work on prosodic guidelines to speech understanding has progressed to a point of considerable success and encouraging results. We have presented theoretical arguments about the need for extracting from the acoustic speech signal some prosodic cues to the large-unit linguistic structure, without dependence upon the prior determination of phonemic structure and recognition of the words in the sentence (Lea, Medress, and Skinner. 1972). Vital assumptions of a prosodically-guided approach to speech understanding have been verified from a variety of experiments. In particular, stressed syllables have been shown to be of prime importance in speech recognition, because of: (a) the occurrence of stressed syllables in semantically important words; (b) the close correspondence between detected phonetic structure and underlying phonemic structures in stressed syllables; (c) the much higher reliability of phonetic classification possible in stressed syllables (as evidenced by the analysis of results from the CMU Speech Segmentation Workshop): and (d) the vital cues to syntactic structure that stressed syllables provide (as evidenced by the different patterns of stress locations for the alternative interpretations of the BBN problem sentences).

A series of "natural experiments" (cf. Anderson, 1966) have been conducted to determine specific relationships between various acoustic prosodic features, on the one hand, and linguistic structures, perceptions, and abstract notions (such as rhythm, etc.), on the other hand. In such "natural experiments", one does not directly control an independent variable (such as syntactic bracketing) and study resultant changes in a dependent variable (such as valleys in F_0 contours); rather, he simply looks at the data obtained from naturally-occurring phenomena (such as the speech previously recorded and identified as the Rainbow Script, Monosyllabic Script, and the 31 ARPA Sentences). From studies of such speech texts, we have demonstrated that over 90% of all intuitively predicted syntactic boundaries are detected from substantial fall-rise valleys in F_0 contours. We have shown that over 85% of all syllables perceived as stressed are located by a particular combination of energy duration, and F_0 cues, assuming archetype F_0 contours for constituents. Available methods for automatic phonetic classifications have been shown to be most reliable within the syllables perceived as stressed, for the 31 ARPA Sentences. Stressed syllabic nuclei have been shown to have a rough tendency toward equal spacing in time, though the dependence upon the number of intervening unstressed syllables is very prominent. Pause durations also appear to correlate with average time intervals between stressed nuclei. Finally, a look at a few similar sentence structures in the BBN problem sentences has shown considerable hope for using prosodic cues to select correct syntactic structures. Those sentences showed distinctive stress patterns, positions of detected constituent boundaries, and F_0 contour parameters (such as the F_0 fall from the peak value in the sentence to the maximum value in the last stressed syllable), which can distinguish yes/no questions and commands, and determine the correct bracketing of the ambiguous word sequences.

In such natural experiments, one cannot be certain that some unknown third variable is not the source of any apparent relationships between the acoustic variable and the underlying abstract variable. Controlled experiments, with all variables except one fixed in the comparison of two utterances, provide the proper extension from the encouraging results of the natural experiments. The designed speech texts provide the necessary controls and sufficient data to extend these encouraging tendencies into well-defined rules relating prosodic variables and linguistic structure.

In the remainder of the current contract period, the design of the controlled speech texts will be completed and they will be recorded by five male and three female talkers, with three repetitions spaced a week or more apart. A subset of those sentences (which subset investigates the effects of stress movement within constituents) will be analyzed. Such analysis includes: obtaining F_0 and energy functions; applying the boundary detection algorithm; locating stressed syllables by algorithm; obtaining listener's stress perceptions for the subset of sentences; conducting a thorough evaluation of the results of these prosodic analyses; and determining whether rules can be written relating stress positions within constituents to expected positions of detected constituent boundaries, acoustic correlates of stress, etc. The algorithm for stressed syllable location is to be implemented on the new Sperry Univac speech research facility.

A report on the designed sentences will be published soon, and a separate description of those sentences that test phonetic sequences (cf. Section 3.3.1) will also be distributed. Also, another report will present a compilation of known rules and testable hypotheses about prosodic patterns and their relationships to speech understanding. This will be used in association with the ARPA workshops on acoustic phonetic and phonological rules. Sperry Univac also intends to continue cooperating with other ARPA contractors on the further developments of speech data bases and the prosodic analysis of selected problem sentences.

In a subsequent contract, we plan to:

(a) Improve and expand the prosodic analysis tools, to provide confidence measures on syntactic boundary locations and stressed syllable locations, and to develop measures of rhythm and rate of speech;

(b) Investigate theoretical predictions of sentence stress from underlying syntactic structure, and experimentally verify such rules with studies of the designed speech texts;

(c) Study prosodic cues to sentence type, syntactic bracketing, subordination, and coordination, using the designed speech texts;

(d) Analyze problem sentences for systems contractors, to see how prosodic information may help select correct structures;

(e) Investigate prosodic cues to word matching, parsing, and semantic analysis in speech understanding systems, using ARPANET access to ARPA systems contractors;

(f) Investigate the combined effects of syntactic, lexical, and phonetic structures on intonation contours, and develop intonation rules that relate fundamental frequency contours to linguistic structures; and

(g) Investigate the use of measures of speech rate and rhythm in phonological analysis procedures of speech understanding systems.

This systematic study will provide some of the most essential tools for using acoustic prosodic data in speech understanding systems. The general strategy for prosodically-guided speech understanding, which was outlined in Section 2, then provides the basic framework within which such knowledge about prosodic regularities can be used to locate reliable phonetic data, select the most likely words for hypothesizing in a sentence, and choose the most promising syntactic and semantic structures to hypothesize and test.

5. REFERENCES

ABE, I. (1967). English Sentence Rhythm and Synchronism, Bull. Phon. Soc. Japan, Vol. 125, 9-11.

ALLEN, G. D. (1967), Two Behavioral Experiments on the Location of the Syllable Beat in Spoken American English, Studies of Language and Language Behavior, vol. 4, 2-179.

ALLEN, G. D. (1968), On Testing for Certain Stress-Timing Effects, Working Papers in Phonetics No. 10, University of California at Los Angeles, 47-59.

ALLEN, G. D. (1972), The Location of Rhythmic Stress Beats in English: An Experimental Study, Language and Speech, vol. 15, 22-100.

ALLEN, J. (1973), Prosodic Contours for Auxiliary Phrases, Paper presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, Calif.

ANDERSON, B F. (1966). The Psychology Experiment. Belmont, Calif.: Brooks/Cole Publishing Co.

ARMSTRONG, L. E. and WARD, I.C. (1926), Handbook of English Intonation. Cambridge; Heffer (2nd Edit.).

BOLINGER, D. A (1958), "A Theory of Pitch Accent in English", Word, vol. 14, p. 109.

BRESNAN, J. (1971), Sentence Stress and Syntactic Transformations, Language, vol. 47, pp. 157-81.

BRESNAN, J. (1972), Stress and Syntax: A Reply, Language, vol. 48, pp. 326-42.

CANTRELL, W. R. (1969), "Pitch, Stress, and Grammatical Relations", Papers from the Fifth Regional Meeting of the Chicago Linguistic Society. Chicago: Univ. of Chicago Press, pp. 12-24.

CHOMSKY, N. (1957), Syntactic Structures. The Hague: Mouton.

CHOMSKY, N. and HALLE, M. (1968), The Sound Pattern of English. New York; Harper and Row.

CHOMSKY, N. and MILLER, G. A. (1963), "Introduction of the Formal Analysis of Natural Languages", in Handbook of Mathematical Psychology, pp. 269-321; Ed. R. D. Luce, R. R. Bush and E. Galanter. New York: John Wiley and Sons, Inc.

FERGUSON, C. F. (1966), Assumptions about Nasals: A Sample Study in Phonological Universals. In Universals of Language (J. H. Greenberg, Ed.). Cambridge, Mass.: MIT Press, 53-60.

HUGHES, G. W., and HEMDAL, J. F. (1965). Speech Analysis Final Report, AF 19-305, Project 5628, Task 562802, for AFCRL, Bedford, Mass., by Purdue Univ. Lafayette, Ind.

HUTTAR, G. L. (1968), Two Functions of the Prosodies in Speech, Phonetica, vol. 18, 231-41.

KLATT, D. H. and STEVENS, K. N. (1972), Sentence Recognition from Visual Examination of Spectrograms and Machine-Aided Lexical Searching, Proc. 1972 Conference on Speech Communication and Processing. IEEE and AFCRL: Bedford, Mass., pp. 315-318.

LEA, W. A. (1972), Intonational Cues to the Constituent Structure and Phonemics of Spoken English, Ph.D. Thesis, School of E.E., Purdue University.

LEA, W. A. (1973a), Syntactic Boundaries and Stress Patterns in Spcken English Texts, Univac Report No. PX 10146, Univac Park, St. Paul, Minnesota.

LEA, W. A. (1973b), An Approach to Syntactic Recognition without Phonemics, IEEE Trans. on Audio and Electroacoustics, vol. AU-21, 249-358.

LEA, W. A. (1973c), Segmental and Suprasegmental Influences on Fundamental Frequency Contours. In Consonant Types and Tone (Proceedings of the First Annual Southern California Round Table in Linguistics, Ed. by L. Hyman), University of Southern California Press.

LEA, W. A (1973d), "Perceived Stress as the 'Standard' for Judging Acoustical Correlates of Stress", Paper presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, Calif.

LEA, W. A (1973e), "Evidence that Stressed Syllables Are the Most Readily Decoded Portions of Continuous Speech", Paper presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, Calif.

LEA, W. A. (1973f), "An Algorithm for Locating Stressed Syllables in Continuous Speech", Paper presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, Calif.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1972a), Prosodic Aids to Speech Recognition: I. Basic Algorithms and Stress Studies, Univac Report No. PX 7940, Univac Park, St. Paul, Minnesota.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1972b), Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition. Presented at the 84th Meeting, Acoustical Society of America, Miami Beach, Florida, No. 27-30, 1972.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1973a), "Prosodic Aids to Speech Recognition: II. Syntactic Segmentation and Stressed Syllable Location", Univac DSD, PX 10232.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1973b), "Prosodic Aids to Speech Recognition: III. Relationships between Stress and Phonemic Recognition Results", Univac DSD, PX 10430.

LEA, W.A., MEDRESS, M.F., and SKINNER, T.E. (1974), A Prosodically-Guided Speech Understanding Strategy, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Penn., 38-44.

LEHISTE, I (1970), Suprasegmentals. Cambridge: M.I.T. Press.

LESSER, V. R , FENNELL, R D , ERMAN, L.D., and REDDY, D.R. (1974), Organization of the Hearsay II Speech Understanding System, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Penn., 11-21.

LIEBERMAN, P. (1967), Intonation, Perception, and Language. Cambridge: M.I.T. Press.

MEDRESS, M F (1969), Computer Recognition of Single-Syllable English Words. Ph.D. Thesis. Dept. of E. E. , M.I.T.

MEDRESS, M (1972), A Procedure for the Machine Recognition of Speech, Proc. 1972 Conf. on Speech Communication and Processing, Newton, Mass., pp. 113-116. April.

MIYAKE, I. (1902), Researches on Rhythmic Action, Studies from the Yale Psychol. Lab., vol. 10, 1-48.

NEWMAN, R., FU, K.S., and LI, K.P. (1972), A Syntactic Approach to the Recognition of Liquids and Glides, Proc. 1972 Conf. on Speech Commun. and Processing. Bedford, Mass.: IEEE and AFCRL, 121-124.

OHALA, J. (1970), Aspects of the Control and Production of Speech, Working Papers in Phonetics No. 15, University of California at Los Angeles.

PIKE, K. L. (1945), The Intonation of American English. Ann Arbor: University of Michigan.

RITEA, H. B. (1974), A Voice-Controlled Data Management System, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Penn., 28-31.

ROVNER, P., MAKHOUL , J., WOLF, J., COLARUSSO, J. (1974), Where the Words Are: Lexical Retrieval in a Speech Understanding System, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Penn., 160-164

SCHWARTZ, R., and MAKHOUL, J. (1974), Where the Phonemes Are: Dealing with Ambiguity in Acoustic-Phonetic Recognition. Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Penn. 85-88.

SHEN, Y., and PETERSON, G.G. (1962), Isochronism in English, Stud. in Ling., Occasional Papers 9, Buffalo: University of Buffalo.

TRAGER, G.L. and SMITH, H.L., JR. (1951), An Outline of English Structure, "Studies in Linguistics: Occasional Papers 3", Norman, Oklahoma; Battenburg Press.

WEEKS, R.V. (1974), Predictive Syllable Mapping in a Continuous Speech Understanding System, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Penn., 154-158.

WOODROW, H. (1951), Time Perception, Handbook of Experimental Psychology (S. S. Stevens, Ed.). New York: Wiley, 1224-1236.

WOODS, W.A. (1974), Motivation and Overview of BBN SPEECHLIS: An Experimental Prototype for Speech Understanding Research, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Penn., 1-10.